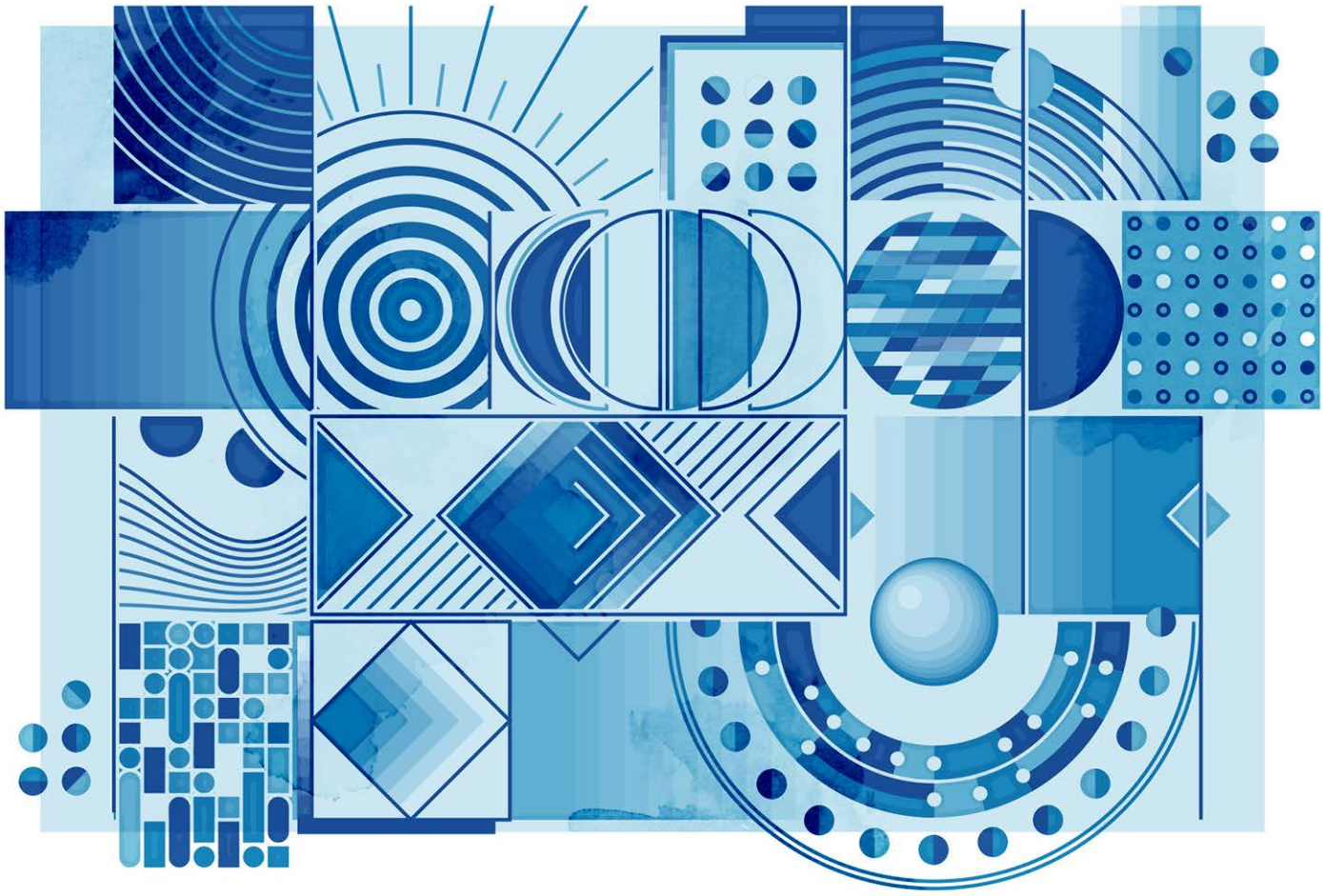
EVALUAR

Revista Evaluar

Laboratorio de Evaluación Psicológica y Educativa

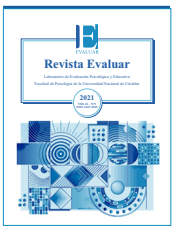
Facultad de Psicología de la Universidad Nacional de Córdoba

2021
VOL 21 - N°3
ISSN 1667-4545



2021, Vol. 21, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



BifactorCalc: An Online Calculator for Ancillary Measures of Bifactor Models

BifactorCalc: Una calculadora en línea para medidas auxiliares de modelos bifactor

José Ventura-León * [1] , Luis Quiroz-Burga [1] ,
Tomás Caycho-Rodríguez [1] , Pablo D. Valencia [2]

1 - Universidad Privada del Norte (UPN), Facultad de Ciencias de la Salud.
2 - Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Iztacala.

Recibido: 01/05/2021 **Revisado:** 15/06/2021 **Aceptado:** 24/06/2021

Abstract

The bifactor model allows examining the presence of a total score in a data set by modeling a general factor and two or more specific factors with an orthogonal relationship. These models tend to overestimate the goodness of fit (e.g., CFI, RMSEA, SRMR), hence there exist auxiliary measures that allow examining the dimensionality (ECV_{Gen}; $ECV_{Specific}$; I-ECV, PUC, ARPB), and reliability (ω, ω_S, ω_H, ω_{HS}, PRV, H, and FD). The present study describes the operation, mathematical foundations, and application in psychological research of an online calculator called *BifactorCalc*. The results demonstrate that *BifactorCalc* is an online, user-friendly, and easy-to-use computer program for the calculation of the different auxiliary measures of bifactor models. It was concluded that the computer tool BifactorCalc is able to calculate the auxiliary measures of bifactor models in three simple steps and generate a *path* diagram.

Keywords: *software, bifactor, SEM, calculator, auxiliary measures*

Resumen

El modelo bifactor permite examinar la presencia de una puntuación total en un conjunto de datos a partir del modelamiento de un factor general y dos o más factores específicos con relación ortogonal. Estos modelos tienden a sobreestimar las bondades de ajuste (v.g., CFI, RMSEA, SRMR), y por esta razón es que existen medidas auxiliares que permiten examinar la dimensionalidad (ECV_{Gen}; $ECV_{Specific}$; I-ECV, PUC, ARPB) y la fiabilidad (ω, ω_S, ω_H, ω_{HS}, PRV, H y FD). El presente estudio describe el funcionamiento, fundamentos matemáticos y aplicación en la investigación psicológica de una calculadora online denominada BifactorCalc. Los resultados demuestran que el *BifactorCalc* es un programa informático online, amigable y de fácil utilización para el cálculo de las diferentes medidas auxiliares de los modelos bifactor. Se concluye que el *BifactorCalc* es una herramienta informática que tiene la capacidad de calcular las medidas auxiliares de modelos bifactor en tres simples pasos y generar un diagrama *path*.

Palabras clave: *software, bifactor, SEM, calculadora, medidas auxiliares*

***Correspondence to:** Dr. José Ventura-León, Facultad de Salud, Universidad Privada del Norte, Av. Alfredo Mendiola, 6062. Los Olivos, Perú. E-mail: jose.ventura@upn.pe

How to cite: Ventura-León, J., Quiroz-Burga, L., Caycho-Rodríguez, T., & Valencia, P. D. (2021). BifactorCalc: An online calculator for auxiliary measures of bifactor models. *Revista Evaluar, 21*(3), 1-14. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar

Participaron en la edición de este artículo: Eva Crasso, Stefano Macri, Juan Balverdi, Alicia Molinari, Mónica Serppe, Eugenia Maiorana, Benjamín Casanova, Ricardo Hernández.

Introduction

In the field of psychological evaluation, whether it be for diagnosis, intervention or research, the objective is to obtain the measurement of the construct (e.g., anxiety, depression, or stress) and the dimensions that comprise it. In this context, it is common to assume that the totality of items is influenced by the same latent variable; composed by specific factors that are integrated in a great general factor (Dominguez-Lara & Rodriguez, 2017). In that sense, the scores of the specific factors are previously summed to obtain a general score. However, it has recently been argued that to conduct this procedure, empirical evidence of the presence of a general factor must be obtained from statistical modeling (Reise, 2012).

In the context of psychology, it is common to use structural equation models (SEM), which are statistical techniques that assume the presence of a latent variable underlying the items and constitute an important alternative for evaluating psychological constructs, which are generally multidimensional (Bonifay, Lane, & Reise, 2017).

Bifactor models (see Figure 1) are within the so-called hierarchical models (Canivez, 2016; Reise, 2012), also called nested factor models (Gustafsson & Balke, 1993), and direct hierarchical models (Gignac, 2008). Its main characteristic is to evaluate the simultaneous effect of a general factor (GF), and specific factors (e.g., F1 and F2), on a set of indicators (Flores-Kanter, Dominguez-Lara, Trógolo, & Medrano, 2018). In that sense, the specific factors (SFs) are assumed to be orthogonal to each other (DeMars, 2013) because the shared variance between the specific factors is due to the general factor (Reise, 2012). Thus, the GF —in comparison to the SFs— is supposed to explain the items' greater amount of variance.

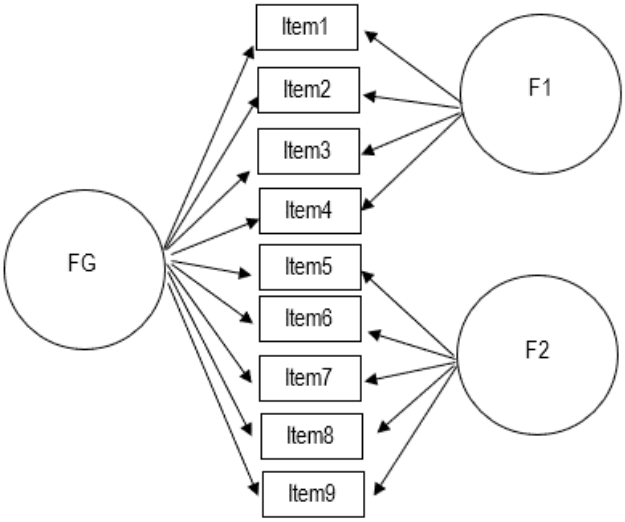


Figure 1
Diagram of bifactor model.

Although the bifactor model was originally described in the late 1930s (Holzinger & Swineford, 1937), it has been rediscovered in the past years (Reise, 2012) and it is increasingly used in psychological research conducted in diverse cultural contexts (Anderson & Marcus, 2019; Montes & Sanchez, 2019; Vuyk & Codas, 2019). Nevertheless, the bifactor model is not immune to criticism. The evaluation of bifactor models with SEM techniques and the traditional goodness-of-fit indices only (e.g., CFI, RMSEA) can lead to false positives since it fails to evaluate the influence of the general factor and specific factors on the items (Bonifay et al., 2017; Dominguez-Lara & Rodriguez, 2017; Flores-Kanter et al., 2018). In fact, the evidence suggests that traditional goodness-of-fit indices may statistically favor bifactor models (Gignac, 2008; Morgan, Hodge, Wells, & Watkins, 2015).

Another important aspect is that the *interchangeability* of the specific factors in symmetric bifactor models (see Figure 1) is a prerequisite for its correct interpretation and the avoidance of anomalous models (see also Eid, Geiser, Koch, & Heene, 2017). The specialized literature provides

Table 1
Comparative information between BifactorCalc and other software.

Comparables	BifactorIndices Calculator	Excel® sheets*	*BifactorCalc*
Installation	R y RStudio	Excel®	None
Skills	R language programming	Using Excel	None
Diagram generation	No	No	Yes. Automatically
Report in APA format	No	No	Calculation output in APA format

Note. *Excel ® developed by Dueber (2017).

some examples of the correct use of the symmetric and *structurally different* bifactor models for the Beck Depression Inventory-II (Heinrich, Zagorscak, Eid, & Knaevelsrud, 2018) and ADHD/ODD symptoms (Burns, Geiser, Servera, Becker, & Beauchaine, 2019).

In this context, it is necessary to have a set of auxiliary measures that allow for a better evaluation of the bifactor model. Specifically software, which is needed to calculate all these measures quickly and easily. Currently, Excel® sheets are available (Dueber, 2017) and an R package called "BifactorIndicesCalculator" (Dueber, 2020) has recently been made available. The latter requires programming skills that are still not common among psychology professionals (comparative information in Table 1). This increases the need for develop a computer program for the calculation of the auxiliary measures of bifactor models, which provides a diagram with the factorial loads entries. In this sense, the objective of this research is to develop a software called *BifactorCalc* that allows for the calculation of the auxiliary measures of bifactor models in an easy, friendly way.

Omega Coefficients

For a bifactor structure, four types of omega coefficients can be calculated: Total Omega (ω), Subscale Omega (ω_S), Hierarchical Omega (ω_H) and Hierarchical Omega for Subscale (ω_{HS}).

The omega coefficient (ω, McDonald, 2013) estimates what proportion of variance in the total observed score can be attributed to all common sources of variance (Reise, Bonifay, & Haviland, 2013). The ω is based on the factor loadings of a factorial model. Unlike other coefficients such as alpha, which is based on the assumption of equal loadings (tau-equivalent models), the omega coefficient is appropriate for cases in which the loadings of the items vary (congeneric models), an indication that is supported by several authors (Dunn, Baguley, & Brunsden, 2013; Rodriguez, Reise, & Haviland, 2015). The calculation of omega is as follows:

$$\omega = \frac{\left(\sum \lambda gen\right)^2 + \sum_{k=1}^{k}\left(\sum \lambda grp_k\right)^2}{\left(\sum \lambda gen\right)^2 + \sum_{k=0}^{k}\left(\sum \lambda grp_k\right)^2 + \sum\left(1-h^2\right)} \quad (1)$$

In the formula, the numerator expresses all common sources of variation of the total weighted score, and the denominator represents all common sources of total variance of the score plus the unique variance. High values of ω indicate high multidimensional composite reliability.

In the same way, the omega coefficient can

be calculated for the specific factors (ω_S) from the factor loadings and errors corresponding to each set of items that comprise the subscale. The following formula is used to calculate ω_S, when the variance of the general factor and the specific factors are combined to estimate reliability:

$$\omega S = \frac{(\sum\lambda gen)^2 + (\sum\lambda_{k=1})^2}{(\sum\lambda gen)^2 + (\sum\lambda_{k=1})^2 + \sum(1-h^2)} \quad (2)$$

Both the ω and ω_S coefficients reflect the systematic variation, attributed to various common factors, that affects weighted composite scores (Rodriguez, Reise, & Haviland, 2016). In this context, it is important to determine the relative weight of the different factors that determine the variance of the composite scores. To that end, some alternate indices have been developed: hierarchical omega (ω_H), and hierarchical omega for the subscale (ω_{HS}). Both ω_H and ω_{HS} reflect the variance attributed to a single latent variable (Rodriguez et al., 2015).

Specifically, ω_H estimates the proportion of variance of the total scores that can be attributed to a single general factor and it is calculated by dividing the squared sum of the factor loadings on the general factor by the variance of the total scores (Reise, Moore, & Haviland, 2013).

$$\omega_H = \frac{(\sum\lambda gen)^2}{(\sum\lambda gen)^2 + \sum_{k=1}^{k} (\sum\lambda grp_k)^2 + \sum(1-h^2)} \quad (3)$$

The ω_H is sensitive to the number of items. A greater number of items is associated with an increase in the ω_H, which is also affected by the relative size of the factor load of each item in the general factor, versus the specific factors (Rodriguez et al., 2015). A high ω_H ($\omega_H > .80$) would express that the scores can be considered essentially unidimensional, since *the general factor* is the main source of systematic variance compared to the influence of the *specific factors*.

The calculation of ω_H can be extended to subscales through the calculation of the hierarchical omega (ω_{HS}), which reflects the proportion of systematic variance of a subscale score after separating the variability attributed to the general factor (Reise, Bonifay et al., 2013). Hierarchical omega is calculated from the following formula:

$$\omega HS = \frac{(\sum\lambda grp_k)^2}{(\sum\lambda gen)^2 + (\sum\lambda grp_k)^2 + \sum(1-h^2)} \quad (4)$$

Thus, there are some cut-off points in psychology that can be used as a reference: $\omega_{HS} \geq .30$ is substantial; $.20 \leq \omega_{HS} < .30$ is moderate and $\omega_{HS} < .20$ is low (Smits et al., 2014).

Percentage of Reliable Variance (PRV)

The percentage of reliable variance (PRV) is an indicator based on the logic of the bifactor model because it considers the variance explained by the general factor (Hammer et al., 2018). This index is the ratio of ω_H to ω; and therefore, it can be conceptually understood as the percentage of the total reliability that can be attributed to the reliability of the general factor (Reise, Moore, et al., 2013). See the following equation:

$$PVR = \frac{\omega_H}{\omega} X100\% \quad (5)$$

Thus, some authors propose as a provisional cut-off point a PVR > 50, which would indicate that half of the reliable variation in test score is produced by the general factor (or the specific one, in which case ω_{HS} is replaced by ω_{H} in the numerator; Li, 2015).

Explained Common Variance (ECV)

The explained common variance (ECV) is an indicator of unidimensionality and expresses the proportion of the common variance that can be attributed to the general factor (Reise, Moore, et al., 2013). For its calculation, the factor loadings of the general and specific factors of a bifactor model are used on the following mathematical expression:

$$ECV_{GEN} = \frac{\sum\lambda^2_{GEN}}{\sum\lambda^2_{GEN}+\sum\lambda^2_{grpk}} \quad (6)$$

Where: $\sum\lambda^2_{GEN}$ is the sum of the squared factor loadings of the general factor; $\sum\lambda^2_{grpk}$ is the sum of the squared factor loads of the specific groups. High ECV values, greater than .60, suggest that the common variance among the specific factors is small compared to the general factor; and therefore, that the data would fit an essentially unidimensional model (Reise, Scheines, Widaman, & Haviland, 2013).

For example, it has been observed that, when the ECV is greater than .60, the correlation between the general factor and a criterion variable is not substantially affected if only the general factor is modeled and not the specific factors. In other words, high ECV values indicate that it is possible to use a unidimensional model even if the data fits better with a bifactor model. Other provisional cut-off points suggested by the literature are .70 or .80 (Rodriguez et al., 2016). However, the interpretation of ECV must be done in conjunction with that of the percentage of uncontaminated correlations (PUC), which is described in the following section (Reise, Scheines, et al., 2013).

$$ECV_{Specific} = \frac{\sum\lambda^2_{grpk}}{\sum\lambda^2_{GEN}+(\sum\lambda grp_k)^2} \quad (7)$$

In the case of the specific factors, a variant of the formula is made by positioning the loadings of the specific factors in the numerator and the loadings of the general factor plus the specific ones in the denominator. On the other hand, it is possible to obtain an ECV for each of the items (ECV-I) with the following mathematical expression:

$$ECV - I = \frac{\lambda^2_{GEN}}{\lambda^2_{grp_n}+\lambda^2_{GEN}} \quad (8)$$

ECV-I expresses the proportion of true variance of each item that is explained by the general factor (Stucky et al., 2013). Values greater than .85 suggest an influence of the general factor on the variance of the item (Stucky & Edelen, 2014).

Percentage of Uncontaminated Correlations (PUC)

The percentage of uncontaminated correlations (PUC; Reise, Scheines, et al., 2013) expresses in percent the amount of correlations that are

not corrupted by multidimensionality (Rodriguez et al., 2015). In other words, it expresses what percentage of the total correlations between items occurs between items belonging to different specific factors. Therefore, the PUC together with the ECV provide information about the bias towards forcing multidimensional data into unidimensional models. Its mathematical expression is presented below:

$$PUC = \frac{\frac{I_G{}^*(I_G-1)}{2} - \left[\frac{I_{S1}{}^*(I_{S1}-1)}{2} + \frac{I_{S2}{}^*(I_{S2}-1)}{2} + \ldots \frac{I_{Sn}{}^*(I_{Sn}-1)}{2}\right]}{\frac{I_G{}^*(I_G-1)}{2}} \quad (9)$$

Where: I_I is the number of items loaded onto the general factor; I_{S1} is the number of items loaded onto the specific factor 1; I_{S2} is the number of items loaded onto the specific factor 2; I_{Sn} is the number of items loaded onto the specific factor *n*.

The interpretation of the PUC must be conducted in conjunction with the ECV. In practical terms, it has been suggested that when the PUC is greater than .80, the ECV value is not very relevant; on the other hand, when the PUC is less than .80, the ECV *should be* greater than .60 in order to treat the instrument as if it were unidimensional (Reise et al., 2013). From another perspective, it has been suggested that when ECV and PUC are both greater than .70, the scale can be treated as if it were unidimensional (Rodriguez et al., 2015).

Factor Determinacy (FD)

Often, researchers do not only model a latent variable, but also seek to estimate everyone's score on that latent variable. These individual scores are called factor scores and, in their simplest form, they correspond to the sum of the items belonging to a factor (DiStefano, Zhu, & Mîndrilă, 2009). There are, however, more refined methods, which are based on estimates from factor analysis.

Nevertheless, one problem with factor scores is that of the so-called indeterminacy. Although the details are technically complex, in simple terms this refers to the fact that from the same factorial solution, it is possible to obtain very dissimilar and even contradictory factor scores. In this sense, FD expresses the multiple correlations between the observed variables (items) and the factor (Grice, 2001). This value can be obtained with the following formula (Beauducel, 2011):

$$FD = diag\left(\Phi\Lambda^T\Sigma^{-1}\Lambda\Phi\right)^{\frac{1}{2}} \quad (10)$$

Under the conditions described in this work, this value is also equivalent to the correlation between factor scores and factors (Beauducel, 2011; Grice, 2001). For this reason, FD varies from 0 to 1, and values close to 1 indicate a better determinacy. In that case, values higher than .80 have been suggested to allow an estimate of the general factor score (Gorsuch, 1983). However, other authors argue for a higher cut-off point (> .90; Grice, 2001; Rodriguez et al., 2015)

Construct Replicability (H)

Another index that can help to better understand the quality of the measurement model is the construct replicability (Mueller & Hancock, 2008). The H index can be used to assess whether the set of items representing a latent variable is adequate. Therefore, it determines if the SEM model is adequate and replicable in all studies. The H index is calculated from the following mathematical formula:

$$H = \frac{1}{\left[1+\frac{1}{\sum_{i=1}^{k}\frac{\lambda_i^2}{1-\lambda_i^2}}\right]} \quad (11)$$

As per the formula above, H is a function of the sum of the factor loading ratios of the squared items (proportion of variance explained by the latent variable), in a factor divided by 1, minus the factor loading squared (Rodriguez et al., 2015). In this sense, as the number of items and the factor loading increase, the H index approaches 1. Values of H greater than .70 suggest that the latent variable is well defined and is more likely to be stable in other studies (Dominguez-Lara, 2016); while low values suggest a poorly defined latent variable, which changes in other studies. The ease of calculating and interpreting the H index, makes it an ideal means of judging the viability of a measurement model based on a set of items.

Average Relative Parameter Bias (ARPB)

The ARPB is a measure for examining the difference between the factor loading of a unidimensional model and the general factor loading of the bifactor model (see equation 12). According to some authors, a maximum difference of .12 to .15 may be acceptable (Rodriguez et al., 2015).

$$ARPB = \sum \frac{\lambda_{GEN} - \lambda_{UNI}}{\lambda_{GEN}} \quad (12)$$

Average Factor Loading (mean)

A first approach to the bifactor model consists in the simple inspection of its factor loadings (Reise, Moore, & Haviland, 2010). If a scale has a strong general factor and a weak set of specific factors, then the factor loadings of the latter will be notoriously low, while the general factor loadings will tend to be higher. A simple way to examine this is by calculating the arithmetic mean of the items. Following other authors, means lower than .30 in the specific factors can be considered secondary evidence of unidimensionality (Ferrando & Lorenzo-Seva, 2017).

Software development

Description of BifactorCalc

The *BifactorCalc* calculator was developed with Python programming language, and all the previously presented formulas were entered. For this purpose, the summation and matrix multiplication calculations were produced with the Numpy library (Harris et al., 2020). The Django framework was used to deploy the web project and build a user-friendly interface in an online version without the need to install the software on a computer (Django Software Foundation, 2019). The styles of the online interface were made with the Bootstrap web style framework, which provided the visual characteristics of the buttons, colors, and frames; tables in APA format, the distribution of the content on the screen, and all the other components displayed on the interface. Finally, the graphic construction of the BifactorCalc and the integration with the calculations were performed with JavaScript.

For use, *BifactorCalc* will require the user and password (Figure 2), which can be requested to the authors of the article via email to store their

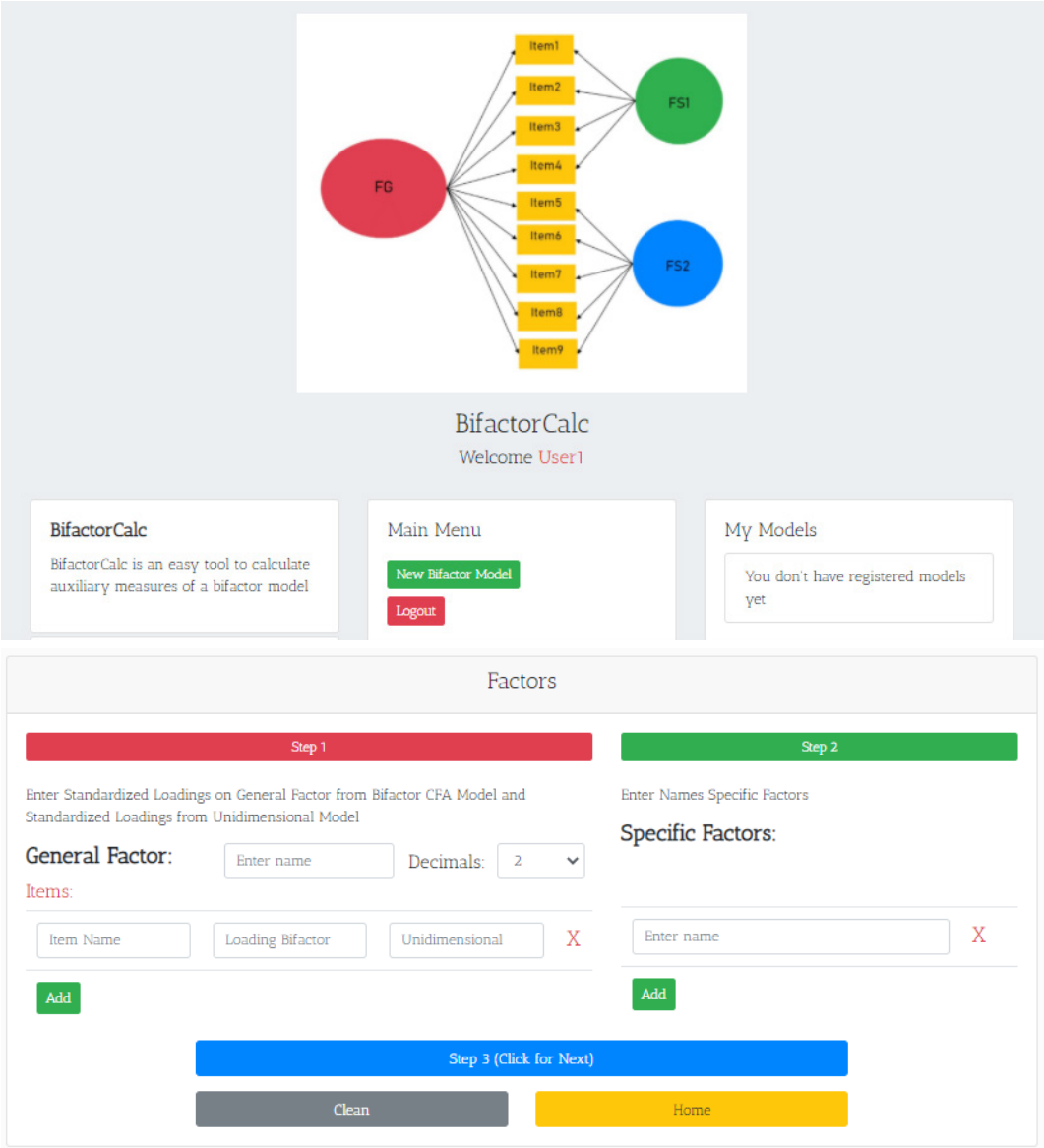


Figure 2
BifactorCalc Welcome Screen and information entry site.

bifactor models privately. Link to calculator: https://joseventuraleon.com/f/bifactorcalc

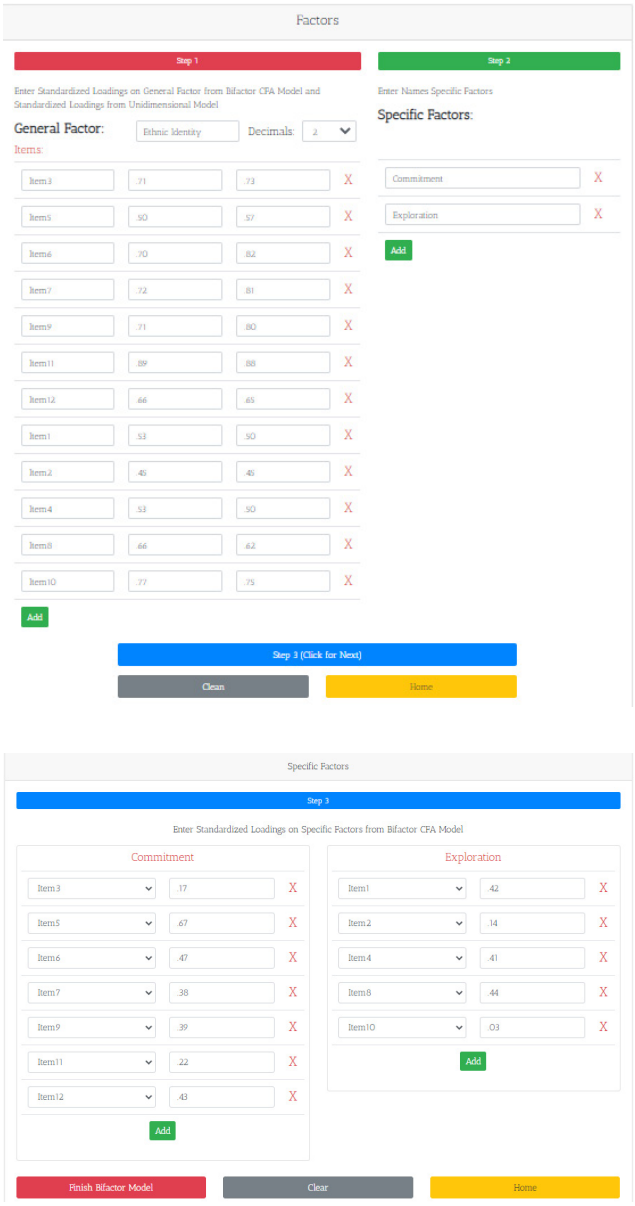
In the main menu there are two options (Figure 2): *New Bifactor Model*, to generate new models and *Logout* to exit the application. In the *My Models* section, the models entered by the user will be displayed, identified by the name of the general factor assigned.

In the *New Bifactor Model* option you can enter the factor loads of the model following the instructions provided in steps 1, 2 and 3. The procedure must be followed so that the calculator receives the data correctly and the information can be calculated satisfactorily.

In Step 1, you must enter the name of the General Factor, the items, and the general factor loadings. In addition, *BifactorCalc* allows for the entry of factor loadings from a unidimensional

model to calculate the ARPB, which measures the difference in the general factor loadings of the bifactor and unidimensional models. In Step 2, you must enter the names of the specific factors of the bifactor model. A maximum of two decimals should be used to enter the loads. Finally, it is necessary to *click* on Step 3, to continue entering the information.

Figure 3
First steps of BifactorCalc.



Fourth, pressing the *Finish Bifactor Model* button automatically performs the calculation of the auxiliary measurements, which appear on a new window, as shown in Figure 3.

Besides, *BifactorCalc* provides a diagram (see Figure 4) that can be copied and used in the scientific manuscript of the *BifactorCalc* user.

Validation of BifactorCalc

To demonstrate the operation of the calculator, information from Yap et al. (2014), was used as a sample, and the similarity of the results obtained in the *BifactorCalc* with the calculations of Rodriguez et al. (2016) was corroborated. Firstly, the factor loadings of the bifactor model provided by Yap et al. (2014) for its Ethnic Identity Scale (EIS) were entered in addition, unidimensional loadings were estimated from the inter-item correlation matrix with the R program (R Core Team, 2020). Secondly, the respective names were assigned to the specific and general factors (see Figure 3 on the left side). Thirdly, the factor loadings of the specific factors were entered (see Figure 3 on the right side).

As for the validation of *BifactorCalc*, the example explaining the "BifactorIndicesCalculator" package was run in R (Deber, 2020) and in *BifactorCalc*. The compared results similarity validate the correct functioning of the software.

Reporting BifactorCalc Results

In relation to the report of a bifactor model, this can be divided into two main moments: (a) Dimensionality, which consists of using the indexes ECV_{Gen}; $ECV_{Specific}$; I-ECV, PUC and ARPB —to determine if the model is unidimensional or multidimensional—; and (b) Reliability, which consists of using the indexes ω, ω_S, ω_H, ω_{HS}, PRV, H and general and specific FD.

Using the information in the example, the

Model

General Factor	N° Items	N° Specific Factors	Decimals	PUC	ARPB
Ethnic Identity	12	2	2	0.53	0.07

General Factor

Name	Omega	OmegaH	PRV	H	FD	ECV	Mean
Ethnic Identity	0.93	0.81	0.87	0.92	0.94	0.75	0.65

Specific Factors

Name	OmegaS	OmegaHS	PRV	H	FD	ECV	Mean
Commitment	0.93	0.22	0.24	0.64	0.79	0.26	0.39
Exploration	0.80	0.16	0.19	0.40	0.69	0.24	0.29

Items

Name	Specific Factor	General Loading	Specific Loading	Standardized Residual Variance	Comunality	ECV-I
Item3	Commitment	0.71	0.17	0.47	0.53	0.95
Item5	Commitment	0.50	0.67	0.30	0.70	0.36
Item6	Commitment	0.70	0.47	0.29	0.71	0.69
Item7	Commitment	0.72	0.38	0.34	0.66	0.78
Item9	Commitment	0.71	0.39	0.34	0.66	0.77
Item11	Commitment	0.89	0.22	0.16	0.84	0.94
Item12	Commitment	0.66	0.43	0.38	0.62	0.70
Item1	Exploration	0.53	0.42	0.54	0.46	0.61
Item2	Exploration	0.45	0.14	0.78	0.22	0.91
Item4	Exploration	0.53	0.41	0.55	0.45	0.63
Item8	Exploration	0.66	0.44	0.37	0.63	0.69
Item10	Exploration	0.77	0.03	0.41	0.59	1.00

Figure 4
Output of the results obtained with BifactorCalc.

Figure 5
Path Diagram of Bifactor Model.

following can be reported: In relation to the dimensionality of the Ethnic Identity Scale, it was observed that it presents an ECV_{Gen} .75, which suggests that the general factor explains 75% of the variance of the items, which could suggest a tendency towards unidimensionality (ECV > .60). In addition, the $ECV_{Specific1}$ and $ECV_{Specific2}$ presented a value of .26 and .24 respectively, which would indicate that the specific factor explains 26% and 24% of the common variance, respectively. In relation to the I-ECV it was observed that only items 3, 11, 2 and 10 are strongly influenced by the general factor (I-ECV > .85). The PUC was equal to .53. Therefore, 53% of the correlations are "contaminated" by the multidimensionality, leaving 47% of the correlations to be explained by the general factor alone. Finally, the ARPB is equal to .07, which indicates that the general factor loads of the bifactor model and factor loads of the unidimensional model are different only by 7%, being within the acceptable ranges.

In relation to the reliability of the IEE, it presented a ω of .93 and ω_S were .93 and .80 for the specific factor 1 and 2 respectively. All these values reveal an excellent composite reliability [the expressions suggested by Cicchetti (1994) used for Cronbach's alpha are extrapolated]. With respect to the ω_H it is equal to .81 expressing that the general factor is the main source of variance in comparison with the specific factors. In this regard, ω_{HS} is .22, which can be considered a moderate consistency of factor 1; and .16 a low consistency of factor 2 (Smits et al., 2014). The PRV would indicate that 87% of the reliable variance is due to the general factor and only 24% and 19% of the reliable variance to the specific factors. The H coefficient is equal to .92 in the general factor, which implies stability in other studies; while the specific Hs are less than .70, providing evidence in favor of the general factor. Finally, the FD for the general factor and the two specific factors are: .94, .79 and .69 respectively, indicating that only the general factor score should be used for the analysis.

Conclusions

This work was aimed at designing a user-friendly, online calculator for the auxiliary measures of the bifactor model. Understanding that multidimensional models are increasingly common (Montes & Sanchez, 2019; Vuyk & Codas, 2019), and the presence of a general factor should be verified empirically (Dominguez-Lara & Rodriguez, 2017; Flores-Kanter et al., 2018). Since the assumption that a high correlation between factors indicates the presence of a total score is no longer sufficient, it is necessary to examine this structure with a bifactor model (Anderson & Marcus, 2019). Some authors state that the bifactor model's goodness-of-fit tends to be positively biased (Bonifay et al., 2017; Gignac, 2008; Morgan et al., 2015) and thus, it is necessary to explore auxiliary measures (Reise et al., 2013; Rodriguez et al., 2016). Despite this, there is no software for the estimation of these measures in a quick and simple way (only in three steps). The most similar option is an R-package (Dueber, 2020), which requires programming skills and the installation of an Office program.

In that sense, *BifactorCalc* is an online software that through a user and password enables the storing of Bifactor models, the modification of factor loadings in case of errors, and the estimation of all auxiliary measures in only three steps. In relation to its validity, information from Yap et al. (2014), and estimates made by Rodriguez et al. (2016), were used to verify that *BifactorCalc*, which reported the same results. This same procedure was performed with

R-package *BifactorIndicesCalculator* (Dueber, 2020). Thus, *BifactorCalc* operation proved to be optimal.

In addition, this research provides an example of how the results obtained with *BifactorCalc* can be reported; framed in two major moments, the review of the dimensionality and the reliability. In this way, the users of this software will be able to easily incorporate the results in their scientific manuscript.

Finally, the software is expected to contribute to the scientific community in the field of psychology and to promote methodological best practices associated with the implementation of the bifactor models in the Spanish-speaking context.

References

Anderson, A. E., & Marcus, D. K. (2019). A bifactor model of meanness, coldheartedness, callousness, and sadism. *Personality and Individual Differences, 137*, 192-197. doi: 10.1016/j.paid.2018.09.006

Beauducel, A. (2011). Indeterminacy of factor score estimates in slightly misspecified confirmatory factor models. *Journal of Modern Applied Statistical Methods, 10*(2), 583-598. doi: 10.22237/jmasm/1320120900

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5*(1), 184-186. doi: 10.1177/2167702616657069

Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2019). Application of the bifactor S-1 model to multisource ratings of ADHD/ODD symptoms: An appropriate bifactor model for symptom ratings. *Journal of Abnormal Child Psychology, 48*(7), 881-894. doi: 10.1007/s10802-019-00608-4

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and Methods of Test Construction: Standards and Recent Advancements* (pp. 247-271). Gottingen, Germany: Hogrefe Publishers.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13*(4), 354-378. doi: 10.1080/15305058.2013.799067

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research and Evaluation, 14*. Retrieved from https://scholarworks.umass.edu/pare

Django Software Foundation. (2019). *Django*. Retrieved from https://djangoproject.com

Dominguez-Lara, S. (2016). Evaluación de la confiabilidad del constructo mediante el Coeficiente H: Breve revisión conceptual y aplicaciones. *Psychologia. Avances de la Disciplina, 10*(2), 87-94. doi: 10.21500/19002386.2134

Dominguez-Lara, S., & Rodriguez, A. (2017). Índices estadísticos de modelos bifactor. *Interacciones. Revista de Avances en Psicología, 3*(2), 59-65. doi: 10.24016/2017.v3n2.51

Dueber, D. M. (2017). *Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models*. doi: 10.13023/edp.tool.01

Dueber, D. M. (2020). Package 'BifactorIndicesCalculator'. Retrieved from https://github.com/ddueber/BifactorIndicesCalculator

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. doi: 10.1111/bjop.12046

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017).

Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22*(3), 541-562. doi: 10.1037/met0000083

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*(5), 762-780. doi: 10.1177/0013164417719308

Flores-Kanter, P. E., Dominguez-Lara, S., Trógolo, M. A., & Medrano, L. A. (2018). Best practices in the use of bifactor models: Conceptual grounds, fit indices and complementary indicators. *Revista Evaluar, 18*(3). doi: 10.35670/1667-4545.v18.n3.22221

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science, 50*(1), 21-43.

Gorsuch, R. L. (1983). Two-and three-mode factor analysis. In R. L. Gorsuch (Ed.), *Factor Analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450. doi: 10.1037/1082-989x.6.4.430

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 2*8(4), 407-434. doi: 10.1207/s15327906mbr2804_2

Hammer, J. H., McDermott, R. C., Levant, R. F., & McKelvey, D. K. (2018). Dimensionality, reliability, and validity of the Gender-Role Conflict Scale-Short Form (GRCS-SF). *Psychology of Men & Masculinity, 19*(4), 570-583.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature 585*, 357-362. doi: 10.1038/s41586-020-2649-2

Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2018). Giving G a meaning: An application of the Bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck Depression Inventory-II. *Assessment, 27*(7), 1429-1447. doi: 10.1177/1073191118803738

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41-54. doi: 10.1007/bf02287965

Li, C. (2015). *The short Grit scale: A dimensionality analysis* (Tesis de maestría). University of Kentucky, EE.UU. Retrieved from http://uk-nowledge.uky.edu/edp_etds/33

McDonald, R. P. (2013). *Test theory: A unified treatment*. New York: Psychology Press. doi: 10.4324/9781410601087

Montes, S. A., & Sanchez, R. O. (2019). El factor p. ¿La estructura subyacente a la psicopatología? *Revista Evaluar, 19*(3), 20-41. doi: 10.35670/1667-4545.v19.n3.26774

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence, 3*(1), 2-20. doi: 10.3390/jintelligence3010002

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. Osborne (Ed.), *Best Practices in Quantitative Method*s (pp. 488-508). Thousand Oaks, California: Sage.

R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from https://www.R-project.org

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129-140. doi: 10.1080/00223891.2012.725437

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6),

544-559. doi: 10.1080/00223891.2010.496477

Reise, S. P., Moore, T. M., & Haviland, M. G. (2013). Applying unidimensional item response theory models to psychological data. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise & M. C. Rodriguez (Eds.), *APA Handbook of Testing and Assessment in Psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 101-119). American Psychological Association. doi: 10.1037/14047-006

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5-26. doi: 10.1177/0013164412449831

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2015). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223-237. doi: 10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. doi: 10.1037/met0000045

Smits, I. A., Timmerman, M. E., Barelds, D. P., & Meijer, R. R. (2014). The Dutch symptom checklist-90-revised. *European Journal of Psychological Assessment, 31*(4), 263-271.

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (1st ed., pp. 183-206).

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment,* 183-206. Routledge/Taylor & Francis Group.

Vuyk, M. A., & Codas, G. (2019). Validación de la Escala de Esperanza Disposicional para Adultos en Paraguay. *Revista Evaluar, 19*(1), 59-71. doi: 10.35670/1667-4545.v19.n1.23880

Yap, S. C. Y., Donnellan, M. B., Schwartz, S. J., Kim, S. Y., Castillo, L. G., Zamboanga, B. L., … & Vazsonyi, A. T. (2014). Investigating the structure and measurement invariance of the Multigroup Ethnic Identity Measure in a multiethnic sample of college students. *Journal of Counseling Psychology, 61*(3), 437-446. doi: 10.1037/a0036253

2021, Vol. 21, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

The Influence of Item Discrimination on Misclassification of Test Takers

Influencia de la discriminación de los ítems en la clasificación incorrecta de los examinados

R. Emmanuel Trujano * [1]

1 - Dirección de Investigación, Calidad Técnica e Innovación Académica, Centro Nacional de Evaluación para la Educación Superior, Mexico City, Mexico.

Recibido: 30/04/2021 **Revisado:** 05/08/2021 **Aceptado:** 08/08/2021

Abstract

It has been suggested that low discriminating items can be included in a test with a criterion-referenced score interpretation as long as they measure a highly relevant content. However, low item discrimination increases the standard error of measurement, which might increase the expected proportion of misclassified test takers. In order to test it, responses from 2000 test takers to 100 items were simulated, varying item discrimination values and number and location of cut scores, and classification inaccuracy was estimated. Results show that the expected proportion of misclassified test takers increased as item discrimination decreased, and as the cut scores were closer to the mean of the distribution of test takers. Therefore, a test should include as few items with low discrimination values as possible —or even none— in order to reduce the expected proportion of test takers classified into a wrong performance level.

Keywords: *decision accuracy, item discrimination, item response theory, Rudner algorithm, information function*

Resumen

Se ha sugerido que en un examen con interpretación de puntajes basada en criterios se pueden incluir ítems con baja discriminación siempre que midan un contenido muy relevante. Sin embargo, los ítems con baja discriminación aumentan el error estándar de medición, lo que podría aumentar la proporción esperada de examinados mal clasificados. Para probarlo, se simularon las respuestas de 2000 examinados a 100 ítems, variando la discriminación de los ítems, el número y ubicación de los puntos de corte, y se estimó la imprecisión de la clasificación. Los resultados muestran que la proporción esperada de examinados mal clasificados aumentó a medida que disminuyó la discriminación de los ítems y que los puntos de corte se acercaron a la media de la distribución de los examinados. Por lo tanto, un examen debería incluir la menor cantidad posible de ítems con baja discriminación —o incluso ninguno— para reducir la proporción esperada de examinados clasificados en un nivel de desempeño incorrecto.

Palabras clave: *precisión de la decisión, discriminación de los ítems, teoría de respuesta al ítem, algoritmo de Rudner, función de información*

***Correspondence to:** R. Emmanuel Trujano, Dirección de Investigación, Calidad Técnica e Innovación Académica, Centro Nacional de Evaluación para la Educación Superior. Avenida Camino al Desierto de los Leones (Altavista) 37, San Ángel, Álvaro Obregón, C.P. 01000, Mexico City, Mexico. Tel.: (+52) 5528205622. E-mail: rmanute@gmail.com

How to cite: Trujano, R. E. (2021). The Influence of Item Discrimination on Misclassification of Test Takers. *Revista Evaluar*, *21*(3), 15-34. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar

Participaron en la edición de este artículo: Rita Hoyos, Andrea Suárez, Eugenia Barrionuevo, Alicia Molinari, Mónica Serppe, Stefano Macri, Florencia Ruiz, Benjamín Casanova, Ricardo Hernández.

Introduction

When a test taker is assessed using a test with a criterion-referenced score interpretation, its performance is referenced to a previously well-defined set of knowledge, skills, or abilities, congruent with the purpose of the test (Popham, 2014; Richaud de Minzi, 2008). If a cut score is set for this test, guidelines for test assembly within the framework of item response theory (IRT) are available, such as maximizing the test information function (TIF) around the cut score value (Lord, 1980), which is accomplished by selecting items with difficulty parameter estimates close to the cut score value and an item discrimination as high as possible (Luecht, 2016).

Another guideline sometimes suggested is to select items that measure contents judged as highly relevant by subject-matter experts (SMEs), even though all test takers —or none— answer them correctly and, therefore, their discrimination parameter estimates are low (Burton, 2001; Clifford, 2016; Frisbie, 2005; Haladyna, 2016; Popham & Husek, 1969). However, this suggestion should be carefully considered because lower item discrimination decreases TIF; since this is associated with an increase in the standard error of measurement and a subsequent increase in the test takers' expected classification inaccuracy (Cheng, Liu, & Behrens, 2015), the inclusion of items with low discrimination estimates may increase the expected proportion of test takers classified into a wrong performance level. Previous research seems to suggest this is the case (Lathrop & Cheng, 2013; Luecht, 2016; Xing & Hambleton, 2004), so the purpose of this research is to further test if item discrimination influences the expected proportion of misclassified test takers.

A simulation study was conducted in which item discrimination was manipulated and the expected proportion of misclassified test takers was recorded. Dichotomously scored test items were simulated because the multiple-choice item is the most used item type among many testing programs (Haladyna, Rodriguez, & Stevens, 2019), and responses are usually scored as correct or incorrect (Haladyna, 2016). In addition, the number of cut scores and their location relative to the test takers' ability distribution was also manipulated since previous research has shown that these factors influence the classification inaccuracy (Ercikan & Julian, 2002; Lathrop & Cheng, 2013; Lee, 2010; Martineau, 2007; Wyse & Hao, 2012). Finally, responses were simulated using either the one-parameter logistic (1PL) or the two-parameter logistic (2PL) IRT model because TIF values obtained with 1PL model are more constrained due to the discrimination value shared by all items whereas TIF values obtained with 2PL model are less constrained due to the variability of discrimination values across items (see Luecht, 2016). Therefore, the more constrained TIF values from 1PL should derive in more classification inaccuracies for 1PL than for 2PL.

Method

This section describes the simulated conditions and the steps I followed to conduct the simulations.

Test Takers Ability Distribution

Samples of 2000 test takers were drawn from a standard normal distribution with a mean of 0 and a standard deviation of 1, which is the a priori distribution used by programs such as BILOG-MG (see Luecht, 2016), IRTPRO (Paek

& Hang, 2013) or R package ltm (Rizopoulos, 2018). These parameters were fixed across all conditions.

Number and Location of Cut Scores

Simulations were conducted with either one or two cut scores. When one cut score was simulated, it could take one of the following values: -1, 0, and 1. Notice that the value of 0 overlaps with the mean of the test takers' ability distribution.

When two cut scores were simulated, the first was always fixed at -1.5 and the second could take one of the following values: 0, 1.5, and 3. Once again, the value of 0 overlaps with the mean of the test takers' ability distribution.

Item Parameters

Responses to 100 dichotomously scored test items were simulated using either the 1PL or the 2PL IRT model.

One Cut Score. When one cut score was simulated, 100 item difficulty values *b* were drawn from a normal distribution with a standard deviation of 1 and a mean equal to each of the cut score values (-1, 0, and 1). In order to reduce variability in the results associated to variability in item difficulty across conditions, the same item difficulties were used to simulate responses with the 1PL and 2PL models and estimate classification inaccuracy.

For the 1PL model, seven item discrimination values *a* were used in the simulations: .25, .5, .75, 1, 1.5, 2, and 2.5. For the 2PL model, 100 discrimination values were drawn from a lognormal distribution with a standard deviation of 1 and one out of seven means: -2, -1, -.5, 0, 1, 2, and 4, subject to the constraint that $0 \leq a \leq 3$. Discrimination values were selected following the classification suggested by Baker and Kim (2017), and DeMars (2010) for their interpretation.

It is important to point out that, once simulated, item discrimination values were paired manually with item difficulty values in such a way so as to maximize test information around the cut score. Specifically, item difficulties were sorted from lowest to highest, whereas item discriminations were sorted to pair the highest values with the difficulties closest to the cut score values. (This was done manually in order to be certain about the location of the maximum values of TIF, since there was a technical problem trying to accomplish it with code).

Table 1 shows descriptive statistics of the simulated difficulty and discrimination values.

Two Cut Scores. When two cut scores were simulated, 50 item difficulties per each cut score value were drawn from a normal distribution with a mean equal to the cut score values, a standard deviation of .5, and a range equal to the mean ± .75. Specifically, for the first cut score (-1.5), *b* was sampled from N (μ = -1.5, σ = .5, *min* = -2.25, *max* = -.75), and when the second cut score was 0, 1.5, or 3, *b* was sampled in the following fashion:

- cut score = 0, *b* was sampled from N (μ = 0, σ = .5, *min* = -.75, *max* = .75).

- cut score = 1.5, *b* was sampled from N (μ = 1.5, σ = .5, *min* = .75, *max* = 2.25).

- cut score = 3, *b* was sampled from N (μ = 3, σ = .5, *min* = 2.25, *max* = 3.75).

Table 1
Descriptive statistics of item parameters for simulations with one cut score.

Mean of sampled distribution				Percentiles (Median)					
	M	SD	Min	P10	P25	P50	P75	P90	Max
Difficulty *b*									
-1.0	-0.993	0.894	-3.437	-2.068	-1.573	-0.939	-0.471	0.219	1.420
0.0	0.025	0.992	-2.202	-1.176	-0.645	0.006	0.650	1.175	3.024
1.0	0.999	1.084	-1.885	-0.228	0.361	1.019	1.605	2.552	3.378
Discrimination *a*									
-2.0	0.264	0.335	0.017	0.051	0.075	0.154	0.307	0.620	2.460
-1.0	0.563	0.564	0.011	0.116	0.228	0.401	0.664	1.161	2.990
-0.5	0.798	0.639	0.057	0.158	0.281	0.641	1.166	1.722	2.529
0.0	1.101	0.676	0.095	0.305	0.552	1.008	1.573	2.140	2.832
1.0	1.501	0.742	0.126	0.600	0.861	1.507	2.177	2.531	2.985
2.0	1.930	0.771	0.224	0.877	1.310	2.024	2.607	2.837	2.996
4.0	2.308	0.574	0.567	1.444	2.001	2.482	2.728	2.905	2.997

These item difficulties were used to simulate responses with the 1PL and 2PL models and estimate classification inaccuracy in order to reduce variability in the results associated with variability in item difficulty across conditions.

For the 1PL model, the same seven item discrimination values a were used in the simulations: .25, .5, .75, 1, 1.5, 2, and 2.5. For the 2PL model, 50 discrimination values per cut score were drawn again from a lognormal distribution with a standard deviation of 1 and one out of seven means: -2, -1, -.5, 0, 1, 2, and 4, subject to the constraint that $0 \leq a \leq 3$. Once again, the highest discrimination values were manually paired with the item difficulties closest to the cut score values in order to maximize test information around the cut scores: for each cut score, item difficulties were sorted from lowest to highest, whereas item discriminations were sorted to pair the highest values with the difficulties closest to the cut score values. (Again, this was done manually in order to be certain about the location of the maximum values of TIF, since there was a technical problem trying to accomplish it with code).

In order to simulate responses to 100 dichotomously scored test items, the 50 item parameters centred at cut score = -1.5 were combined with the 50 item parameters centred at each of the remaining cut scores. Table 2 shows descriptive statistics of each combination of simulated difficulty and discrimination values.

In summary, a total of 2 (one or two cut scores) × 3 (cut score values) × 2 (IRT models) × 7 (item discrimination values) conditions were simulated. Within each condition, the expected proportion of misclassified test takers was calculated with the Rudner algorithm (Rudner, 2001, 2005), which assumes that an individual's estimated ability $\hat{\theta}$ follows a normal distribution with mean θ and standard error $SE_{(\theta)} = 1/\sqrt{I(\theta)}$ (that is, the inverse of the square root of TIF). If an individual's estimated ability is below the cut score value, the probability of misclassification is the area under the normal distribution which is above the cut score. Conversely, if an individual's estimated ability is above the cut score value, the probability of misclassification is the area under the normal distribution which is below the cut score. The expected proportion of misclassified

Table 2
Descriptive statistics of item parameters for simulations with two cut scores.

Mean of sampled distribution				Percentiles (Median)					
	M	SD	Min	P10	P25	P50	P75	P90	Max
Difficulty *b* [a]									
$\mu_{cs2} = 0.0$	-0.735	0.839	-2.235	-1.898	-1.483	-0.729	-0.045	0.350	0.606
$\mu_{cs2} = 1.5$	0.024	1.555	-2.235	-1.898	-1.483	0.030	1.509	1.805	2.230
$\mu_{cs2} = 3.0$	0.753	2.272	-2.235	-1.898	-1.483	0.800	2.991	3.337	3.629
Discrimination *a*									
-2.0	0.245	0.280	0.005	0.048	0.077	0.154	0.308	0.570	1.874
-1.0	0.592	0.542	0.023	0.137	0.237	0.410	0.745	1.267	2.772
-0.5	0.713	0.493	0.041	0.159	0.340	0.633	1.022	1.467	2.302
0.0	1.136	0.688	0.105	0.334	0.599	1.048	1.573	2.182	2.922
1.0	1.557	0.736	0.229	0.598	0.961	1.521	2.111	2.613	2.876
2.0	1.974	0.634	0.526	1.141	1.436	2.027	2.519	2.792	2.982
4.0	2.296	0.631	0.410	1.390	1.961	2.494	2.821	2.932	2.997

Note. [a]The table shows descriptive statistics of 100 item difficulties sampled from two normal distributions: 50 from a distribution with a mean of $\mu_{cs1} = -1.5$ (the first cut score), and 50 from a distribution with a mean equal to each value of the second cut score (μ_{cs2}).

test takers is the average across individuals of the probabilities of misclassification.

Data Generation Steps

Within each condition, data were generated as follows:

1. Set the number of cut scores.
2. Set the cut score values.
3. Set the item parameter values.
4. Draw a sample of 2000 test takers from a standard normal distribution.
5. Simulate 100 responses to dichotomously scored test items with either the 1PL or the 2PL IRT model.
6. Estimate the test takers' maximum likelihood ability and their standard error of measurement according to their simulated responses and the item parameter values.
7. For each test taker, estimate the probability of misclassification, that is, the probability of being classified into performance level B when its estimated ability level falls into performance level A (p(B|A)).
8. Estimate the overall expected classification inaccuracy by averaging all the individual probabilities of misclassifications.
9. Estimate the expected proportion of false positives and false negatives when a single cut score was simulated, or the expected proportion of each misclassification when two cut scores were simulated, by averaging the corresponding individual probabilities of misclassifications.

10. Repeat steps 4 to 9 1000 times.

All the simulations were conducted in the R statistical software (R Core Team, 2020): Item parameters were simulated with package Runuran (Leydold & Hörmann, 2021); responses to items were simulated and test takers' abilities were estimated with package irtoys (Partchev, Maris, & Hattori, 2017); data were plotted with package ggplot2 (Wickham, 2016; Wickham et al., 2021); and the expected proportions of misclassified test takers were estimated with code adapted from package cacIRT (Lathrop, 2014, 2015). Appendix 1 shows the code used to conduct simulations with one cut score, whereas Appendix 2 shows the code for simulations with two cut scores.

Results

Figure 1 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 1PL IRT model. The general trend across all panels is that the expected misclassification values decrease as item discrimina-

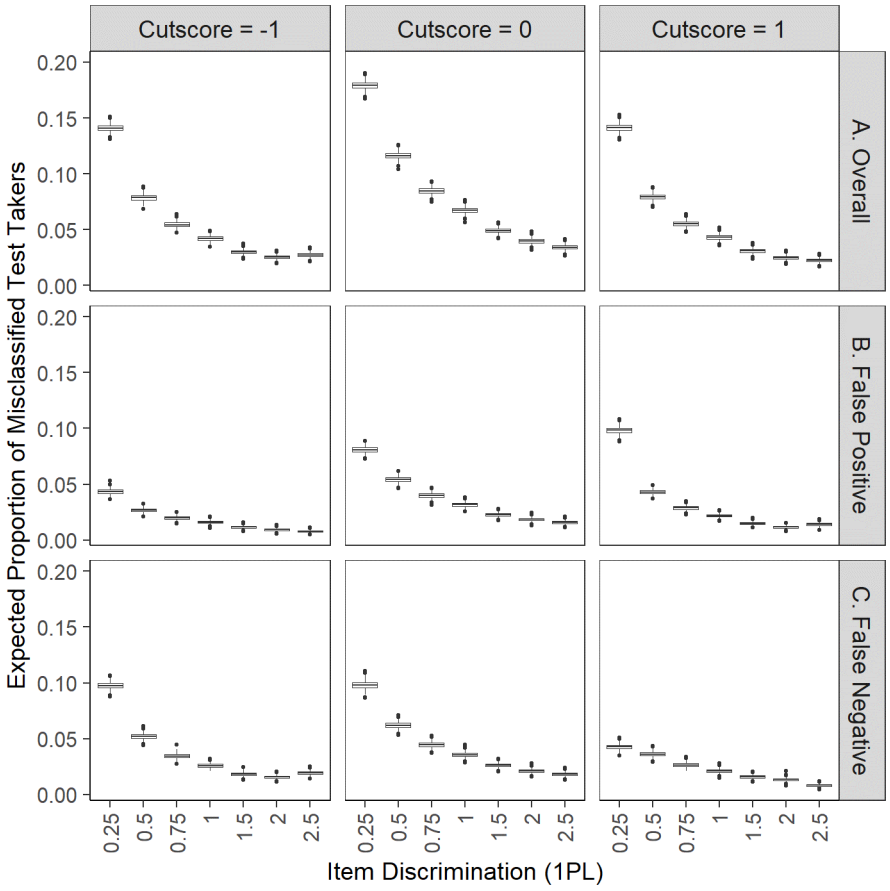


Figure 1
Expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 1PL IRT model.

Note. Data points represent outliers within each condition.

tion increases.

The top panels show that the overall expected classification inaccuracies are higher when the cut score is placed in the mean of the test takers' ability distribution, irrespective of item discrimination values, and lower otherwise. Compare, for example, the boxplots of overall misclassifications in the top panels when item discrimination equals .25.

With one cut score, misclassifications are of two types: a false positive (being wrongly classified into the higher performance level when estimated ability score belongs to the lower one) and a false negative (being wrongly classified into the lower performance level when estimated ability score belongs to the higher one). The middle and bottom panels of Figure 1 show expected misclassification values separated into false positives and negatives, respectively. Within each item discrimination value, false positives are comparable when cut score equals 0 and 1, and lower when cut score equals -1; as an example, compare the boxplots of false positives in the middle panels when item discrimination equals .25. In contrast, false negatives are comparable for each item discrimination value when cut score equals -1 and 0, and lower when cut score equals 1; for example, compare the boxplots of false negatives in the bottom panels when item discrimination equals .25.

Notice that all the distributions are narrow, and even the data points corresponding to outliers at each boxplot are not far away between each other. This implies that the expected misclassification values are consistent within each condition.

Figure 2 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 2PL IRT model. The same trends as those using 1PL model can be observed: the expected misclassification values decrease as the median of item discrimination increases across all panels, the overall expected classification inaccuracies when the cut score is placed in the mean of the test takers' ability distribution are higher than when it is placed farther (irrespective of the median of item discrimination values), false positives are lower when cut score equals -1 than when it equals 0 and 1, false negatives are lower when cut score equals 1 than when it equals -1 and 0, and all the distributions are narrow (implying that expected misclassification values are consistent within each condition).

When Figures 1 and 2 are compared, the proportions of misclassified test takers are lower for the 2PL than for the 1PL model. As an instance, compare the top panels of both figures: the maximum overall expected misclassifications reach a median of .15 for 2PL when the cut score equals 0, lower than the median of approximately .18 for 1PL.

Figure 3 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 1PL IRT model. Notice the different scales on the *y*-axis. Consistent with the results of simulations with one cut score, the expected misclassification values decrease as item discrimination increases.

The top panels show that the overall expected classification inaccuracies decrease as the value of the second cut score shifts away from the mean of the test takers' ability distribution, but lower item discrimination is still associated to higher expected misclassification of test takers.

With two cut scores and three performance levels, there are two misclassification types per each performance level: a false performance level 2 and a false performance level 3 when estimated ability score belongs to performance level 1, a false performance level 1 and a false performance

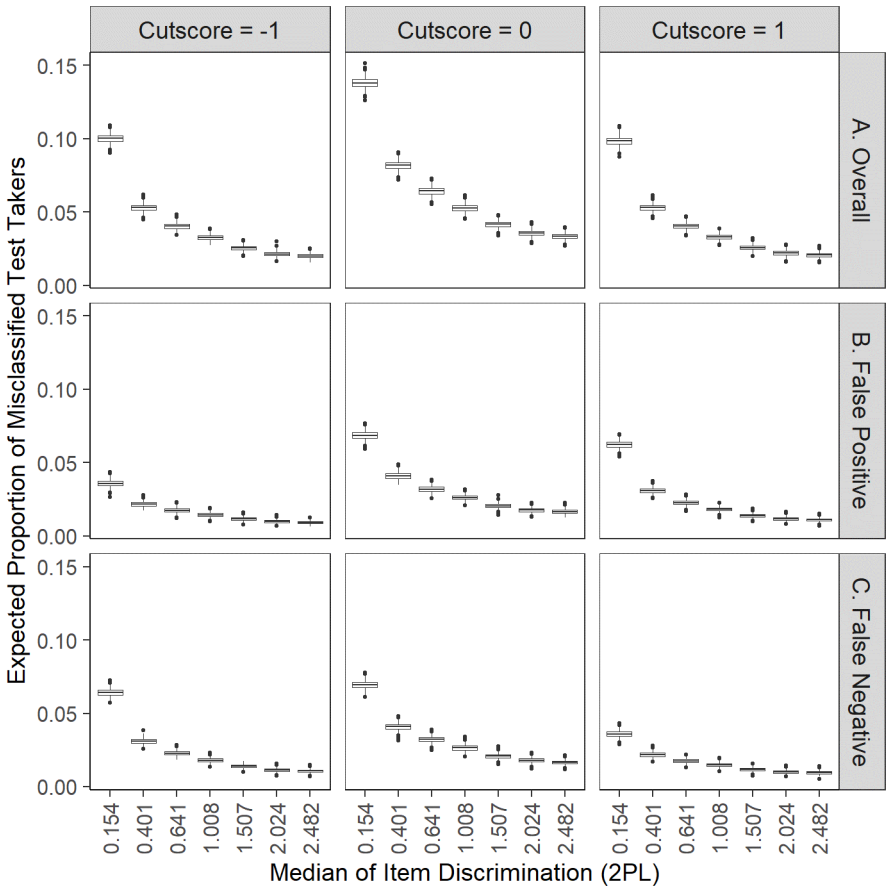


Figure 2
Expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 2PL IRT model.

Note. Data points represent outliers within each condition.

level 3 when estimated ability score belongs to performance level 2, and a false performance level 1 and a false performance level 2 when estimated ability score belongs to performance level 3. The remaining rows of Figure 3 show these six expected misclassification values.

The second row of Figure 3 shows that item discrimination is associated to a decreasing expected false performance level 2 classification of test takers whose estimated ability score belongs to performance level 1, and this trend is comparable across the cut score values. On the other hand, the third row shows that it is unlikely for test takers to be misclassified into performance level 3 if their estimated ability score belongs to performance level 1, unless the second cut score is placed in the mean of ability distribution and item discrimination is as low as .25. Although this is consistent with the general trend observed until now, misclassification into performance level 3 is still unlikely.

The fourth and fifth rows of Figure 3 show that increasing item discrimination is again associated to a decreasing expected misclassification of test takers into performance levels 1 and 3 when their estimated ability score belongs to performance level 2. In the first case, the false performance level 1 classification is comparable across cut score values because the first cut score was always fixed, so the proportion of misclassi-

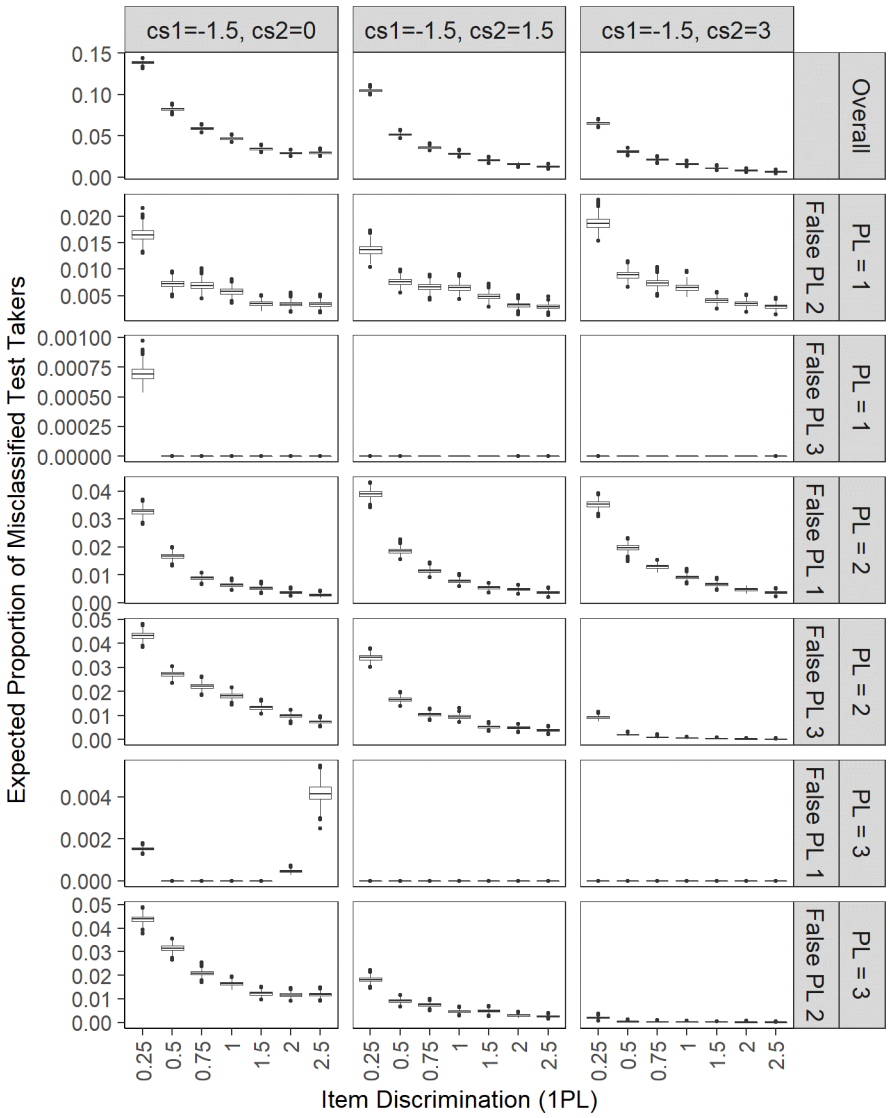


Figure 3

Expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 1PL IRT model.

Note. Data points represent outliers within each condition. Notice the different scales in the *y*-axis. cs1 = cut score 1 (always fixed at -1.5); cs2 = cut score 2; PL = performance level.

fied test takers remained somewhat stable across simulations. In the second case, the false performance level 3 classification decreased as the second cut score shifted away from the mean of ability distribution, but misclassifications still increased as item discrimination decreased.

The last two rows of Figure 3 show the expected misclassification of test takers into performance levels 1 and 2 when their estimated ability score belongs to performance level 3. The sixth row shows that it is unlikely for test takers to be misclassified into performance level 1 if their estimated ability score belongs to performance level 3. An exception occurred when the second cut score is placed in the mean of ability distribution and item discrimination is 2.5, which is inconsistent with the general trend observed previously because one misclassification increased with a high

item discrimination. The reason for this exception is that TIF associated to an item discrimination of 2.5 is lower than TIF associated to other item discrimination values (data not shown) for ability values of 1 or higher, that is, at performance level 3; this lower information increased the standard error of measurement more than for other item discrimination values and increased test takers' expected classification inaccuracy. The last row of Figure 3 shows that the false performance level 2 classification decreased as the second cut score shifted away from the mean of ability distribution, but once again, misclassifications still increased as item discrimination decreased.

Finally, Figure 4 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 2PL IRT model. Notice again the different scales on the *y*-axis.

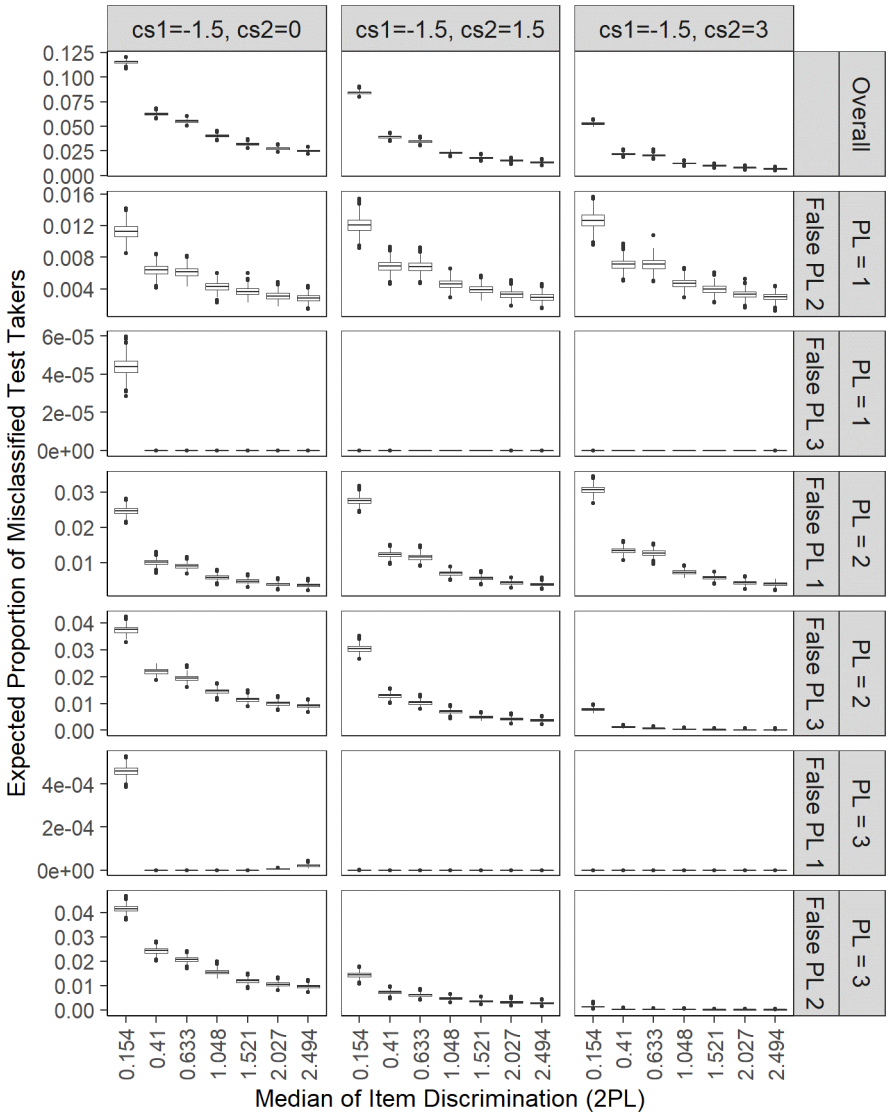


Figure 4
Expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 2PL IRT model.

Note. Data points represent outliers within each condition. Notice the different scales in the *y*-axis. cs1 = cut score 1 (always fixed at -1.5); cs2 = cut score 2; PL = performance level.

Results using 2PL model are similar to those using 1PL model; if anything, the exception found using 1PL model when the second cut score is placed in the mean of ability distribution and item discrimination is 2.5 was not reproduced using 2PL model.

When Figures 3 and 4 are compared, the proportions of misclassified test takers are lower for the 2PL than for the 1PL model. Compare, for example, the top panels of both figures: the maximum overall expected misclassifications reach a median of approximately .112 for 2PL when the second cut score equals 0, lower than the median of approximately .140 for 1PL.

Discussion

The present simulations were conducted in order to test whether item discrimination influences the expected proportion of misclassified test takers. Consistent with previous studies (Lathrop & Cheng, 2013; Luecht, 2016; Xing & Hambleton, 2004), the results suggest this is the case: a test with low discriminating items tends to increase the proportion of misclassified test takers, irrespective of the location of the cut score. Only one exception was observed: Misclassifications into performance level 1 of test taker who should be placed into performance level 3 were higher when two cut scores were simulated, the second cut score was placed in the mean of ability distribution and item discrimination using the 1PL model was 2.5 (see Figure 3). However, this result does not undermine the general conclusion because this increase in misclassifications is associated to a lower TIF for ability values of 1 or higher; since item discrimination and the observed exception are both associated to TIF, which at the same time is associated to a correct -or incorrect- classification of test takers, then the general result and the observed exception do not contradict each other because both support that TIF is associated to classification inaccuracy.

In addition, simulation using the 2PL IRT model yielded less expected classification inaccuracies than simulation using the 1PL model. This may be attributed to the variability in discrimination values simulated under the 2PL model: Even when the median was as low as .154, there was at least one item with high discrimination (see the last column of Tables 1 and 2), thus increasing TIF and, therefore, reducing misclassifications.

The present results also replicated those of previous research suggesting that the number of cut scores and their location relative to the test takers ability distribution influence classification inaccuracy (Ercikan & Julian, 2002; Lathrop & Cheng, 2013; Lee, 2010; Martineau, 2007; Wyse & Hao, 2012). Specifically, misclassifications tended to decrease as the cut score value shifted away from the mean of the ability distribution, and this trend was more notorious when item discriminations (1PL) or their median (2PL) were low. Besides, with one cut score, false positives and negatives follow a different trend depending on the cut score value: the former decreased only when cut score equals -1, whereas the latter decreased only when cut score equals 1. This may have some implication depending on the cost associated to each misclassification (for example, when establishing a cut score to detect intellectual risk; see Ramírez-Benítez, Jiménez-Morales, & Díaz-Bringas, 2015), and minimization of one at the expense of the other may be carefully considered on each particular case. But still, item discrimination may help diminish this problem.

With two cut scores, misclassification into an adjacent performance level was higher than misclassification into a further one. Test takers classified into performance level 1 were more

likely to be misclassified into performance level 2 than into performance level 3; conversely, test takers classified into performance level 3 were more likely to be misclassified into performance level 2 than into performance level 1. In addition, test takers classified into performance level 2 were the most likely to be misclassified into any other performance level than the remaining test takers, but as a consequence of fixing the first cut score at -1.5, they were less likely to be misclassified into performance level 3 as the second cut score shifted away from the mean of the ability distribution, and this shift did not influence misclassification into performance level 1. But once again, item discrimination may help diminish this problem irrespective of the location of cut score values.

The manipulation of the second cut score simulated in the present study may not be far from some real-life situations. As an example, suppose that test takers who get scores at performance level 3 are candidate for an award; if the stakeholders decide to increase the minimum score to be classified into that level, some test takers who could have been eligible for receiving the award will no longer be considered because their test score will now belong to performance level 2. However, this decision will reduce the expected proportion of misclassified test takers in both performance levels. Every particular case should weigh the importance of each consequence in order to make a decision, but once again, increasing item discrimination may help stakeholders in decision making since misclassifications would be a less complicated issue to weigh.

Ercikan and Julian (2002) and Martineau (2007) further showed that the expected classification inaccuracy increases as the number of classification categories increases. Although the present study did not simulate conditions that are fully comparable between one and two cut scores, an exercise can be made only for illustrative purposes: comparing the overall expected proportion of misclassified test takers in conditions where the single cut score equals 0 versus conditions with two cut scores where the second cut score equals 0. This means comparing the top middle panel of Figure 1 with the top left panel of Figure 3, as well as the top middle panel of Figure 2 with the top left panel of Figure 4. Table 3 shows these comparisons more directly by reproducing the median misclassification values of the conditions just mentioned, their interquartile deviation (which is [P75 – P25]/2) and the difference between the medians. Positive differences imply higher misclassification with one cut score, negative differences imply higher misclassification with two cut scores.

As can be seen, all differences suggest that misclassifications were higher with one cut score that with two, which is contrary to the two studies previously mentioned. However, the differences are negligible and they decrease as item discrimination increases. One possible reason for this apparent discrepancy is the number of items simulated: the present study simulated responses to 100 items, whereas the previous studies simulated less than 60, so the present study simulated conditions in which negligible differences could be found since more items tended to increase TIF and, therefore, reduce the standard error of measurement. But once again, these results should be taken only as an illustrative exercise since comparability between conditions is not warranted.

Two limitations of the present results need to be considered. First, this study used the Rudner (2001, 2005) algorithm for estimating expected misclassification of test takers, which assumes that an IRT model fit data appropriately and that ability is normally distributed. This implies that the present results may not be generalizable to cases where these assumptions do not hold. In that

Table 3
Median (and interquartile deviation) of the overall expected proportion of misclassified test takers, and their difference, as a function of either item discrimination (1PL) or median of item discrimination (2PL).

	Expected misclassifications by number of cut scores		
Discrimination	**One (cs = 0)**	**Two (cs2 = 0)**	**Difference**
1PL			
0.250	.179 (0.0024)	.139 (0.0012)	.040
0.500	.116 (0.0022)	.083 (0.0011)	.033
0.750	.085 (0.0020)	.059 (0.0011)	.026
1.000	.067 (0.0018)	.047 (0.0010)	.020
1.500	.049 (0.0016)	.034 (0.0010)	.015
2.000	.040 (0.0015)	.029 (0.0009)	.011
2.500	.034 (0.0014)	.030 (0.0009)	.004
2PL			
0.154	.138 (0.0023)	.116 (0.0011)	.022
0.410	.082 (0.0020)	.063 (0.0010)	.019
0.633	.064 (0.0019)	.055 (0.0010)	.009
1.048	.053 (0.0016)	.040 (0.0010)	.013
1.521	.042 (0.0015)	.032 (0.0009)	.010
2.027	.036 (0.0015)	.028 (0.0008)	.008
2.494	.033 (0.0014)	.025 (0.0008)	.008

case, a nonparametric algorithm such as Lathrop and Cheng's (2014) may be more appropriate to estimate classification inaccuracy, but it is difficult to say whether item discrimination influences inaccuracy since no explicit relation has been stated between these two in an algorithm like that: the probability of a correct response is conditional on observed total score and these two are not related by an item characteristic curve determined by item parameters.

Another limitation is that item parameters were not estimated, instead, the simulated true values were used in estimation of the expected proportion of misclassified test takers, so capitalization on chance was not investigated. Hambleton and Jones (1994) mentioned that item parameter estimates have a positive error relative to their true values, and Yen (1987) showed that this error is bigger for discrimination parameter estimates. For the present study, this implies that TIFs could have been overestimated, and thus misclassifications in general would have decreased, had calibrated item parameter estimates been used. However, there is no reason to suspect that capitalization on chance has a differential influence on item discrimination depending on their true values, so it is possible that the main result of the present study may remain using parameter estimates instead of true values. In addition, Hambleton and Jones (1994) found that a sample size of 2000 test takers, such as the one used in the present study, reduces the effect of capitalization on chance, and Yen (1987) reported a diminishing effect of capitalization on chance as test length increased (see her Table 3), so it is reasonable to suspect that the same would have happened in this study, had calibrated item parameter estimates been used.

In summary, this study found that item discrimination has a negative association with the expected proportion of misclassified test takers: the higher the item discrimination becomes, the lower expected misclassification will be observed. In a test with a criterion-referenced score interpretation, it is important to get validity evidence based on test content (Popham & Husek, 1969), which is the reason that justifies the inclusion of item that don't fully discriminate (Burton, 2001; Clifford, 2016; Frisbie, 2005; Haladyna, 2016; Popham & Husek, 1969). Nevertheless, it is recommended to include as few items with low discrimination values as possible —or even none— because, otherwise, it becomes more likely to classify a test taker into a wrong performance level.

References

Baker, F. B., & Kim, S.-H. (2017). *The basics of Item Response Theory using R.* New York, N.Y: Springer. doi: 10.1007/978-3-319-54205-8

Burton, R. F. (2001). Do item-discrimination indices really help us to improve our tests? *Assessment & Evaluation in Higher Education, 26*(3), 213-220. doi: 10.1080/02602930120052378

Cheng, Y., Liu, C., & Behrens, J. (2015). Standard error of ability estimates and the classification accuracy and consistency of binary decisions. *Psychometrika, 80*(3), 645-664. doi: 10.1007/s11336-014-9407-z

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals, 49*(2), 224-234. doi: 10.1111/flan.12201

DeMars, C. (2010). *Item Response Theory.* Oxford, Oxfordshire: Oxford University Press.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*(3), 269-294. doi: 10.1207/S15324818AME1503_3

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*(3), 21-28. doi: 10.1111/j.1745-3992.2005.00016.x

Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed, pp. 392-409). New York, NY: Routledge.

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are multiple-choice items too fat? *Applied Measurement in Education, 32*(4), 350-364. doi: 10.1080/08957347.2019.1660348

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*(3), 171-186. doi: 10.1207/s15324818ame0703_1

Lathrop, Q. N. (2014). R package cacIRT: Estimation of classification accuracy and consistency under item response theory. *Applied Psychological Measurement, 38*(7), 581-582. doi: 10.1177/0146621614536465

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package cacIRT. *Practical Assessment, Research, and Evaluation, 20*, Article 18. Retrieved from https://scholarworks.umass.edu/pare/vol20/iss1/18

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement, 37*(3), 226-241. doi: 10.1177/0146621612471888

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. *Journal of Educational Measurement, 51*(3), 318-334. doi: 10.1111/jedm.12048

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1-17. doi: 10.1111/j.1745-3984.2009.00096.x

Leydold, J., & H"ormann, W. (2021). Runuran: R interface to the 'UNU.RAN' random variate generators (Version 0.34) [R package]. Retrieved from https://

cran.r-project.org/web/packages/Runuran/index.html

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2016). Applications of item response theory: Item and test information functions for designing and building mastery tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (2nd ed.). New York, NY: Routledge.

Martineau, J. A. (2007). An expansion and practical evaluation of expected classification accuracy. *Applied Psychological Measurement, 31*(3), 181-194. doi: 10.1177/0146621606291557

Paek, I., & Han, K. T. (2013). IRTPRO 2.1 for Windows (item response theory for patient-reported outcomes). *Applied Psychological Measurement, 37*(3), 242-252. doi: 10.1177/0146621612468223

Partchev, I., Maris, G., & Hattori, T. (2017). irtoys: A collection of functions related to item response theory (IRT) (Version 0.2.1) [R package]. Retrieved from https://cran.r-project.org/package=irtoys

Popham, W. J. (2014). Criterion-referenced measurement: Half a century wasted? *Educational Leadership, 71*(6), 62-66. Retrieved from http://www.ascd.org/publications/educational_leadership/mar14/vol71/num06/Criterion-Referenced_Measurement@_Half_a_Century_Wasted%C2%A2.aspx

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1-9. doi: 10.1111/j.1745-3984.1969.tb00654.x

R Core Team. (2020). R: A language and environment for statistical computing (Version 4.0.2). [Computer software]. Retrieved from https://www.R-project.org

Ramírez-Benítez, Y., Jiménez-Morales, R. M., & Díaz-Bringas, M. (2015). Matrices progresivas de Raven: Punto de corte para preescolares 4 - 6 años. *Revista Evaluar, 15*(1), 123-133. doi: 10.35670/1667-4545.v15.n1.14911

Richaud de Minzi, M. C. (2008). Nuevas tendencias en psicometría. *Revista Evaluar, 8*(1), 1-19. doi: 10.35670/1667-4545.v8.n1.501

Rizopoulos, D. (2018). ltm: Latent trait models under IRT (Version 1.1-1) [R package]. Retrieved from https://CRAN.R-project.org/package=ltm

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research, and Evaluation, 7*, Article 14. doi: 10.7275/an9m-2035

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research, and Evaluation, 10*, Article 13. doi: 10.7275/56a5-6b14

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). New York, NY: Springer. doi: 10.1007/978-3-319-24277-4

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., ... & RStudio. (2021). ggplot2: Create elegant data visualisations using the grammar of graphics (Version 3.3.5) [R package]. Retrieved from https://cran.r-project.org/web/packages/ggplot2/index.html

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement, 36*(7), 602-624. doi: 10.1177/0146621612451522

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5-21. doi: 10.1177/0013164403258393

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*(2), 275-291. doi: 10.1007/BF02294241

Appendix 1. R Code Used to Conduct Simulations with One Cut Score

This appendix shows the R code I used to simulate each condition with one cut score. It is important to point out that after step 3a of the code I manually paired item discriminations with item difficulty values in such a way as to maximize test information around the cut score. Once done, the rest of code can be run.

```
#Simulate expected misclassifications with IRT 1PL or 2PL models
#One cut score

#Load necessary R packages
library(tidyverse)
library(irtoys)
library(Runuran)

#Set constants
nsimul <- 1000    #Number of replications of steps 4 to 9 according to step 10
n <- 2000            #Number of test takers according to step 4
k <- 100             #Number of items

#Steps 1 and 2: Set the number and value of cut scores
cutscore <- 0
#Step 3: Set the item parameter values
#3a: Sampling or setting item parameters
#Sampling item difficulty values
b <- rnorm(k,0,1)
#Fixing guessing parameter values at zero
c <- rep(0,k)
#When using 1PL model, set discrimination with these two lines
a1pl <- 0.25       #Item discrimination value
a <- rep(a1pl,k)
#When using 2PL model, sample discrimination with these two lines
meanlog <- -2    #Mean of sampled lognormal distribution
a <- urlnorm(k, meanlog, 1, 0, 3)
#3b: Creating table of item parameter values
#(column 1=a, column 2=b, column 3=c)
params <- cbind(a,b,c)
```

```
params <- data.matrix(params)

#Loop to perform 1000 replications
Data <- rep(NA,nsimul*3)
for (s in 1:nsimul) {
   #Step 4: Draw a sample of 2000 test takers from a
   #            standard normal distribution (~N(m=0,sd=1))
   theta_sim <- rnorm(n,0,1)
   #Step 5: Simulate 100 responses to dichotomously scored test items
   resps <- sim(params,theta_sim)
   #Step 6: Estimate test takers' maximum likelihood ability
   #            and standard error of measurement
   theta_obs <- mlebme(resps,params,method="ML")

   #Step 7: Estimate individual probability of misclassification
   inacc <- matrix(NA,n,2)
   for (i in 1:length(theta_sim)) {
      if (theta_obs[i,1]<cutscore) {
         #If ability < cutscore, label and estimate false positive
         inacc[i,1] <- 0
         inacc[i,2] <- pnorm(Inf,theta_obs[i,1],theta_obs[i,2]) -
                              pnorm(cutscore,theta_obs[i,1],theta_obs[i,2])
         }
      else {
         #If ability >= cutscore, label and estimate false negative
         inacc[i,1] <- 1
         inacc[i,2] <- pnorm(cutscore,theta_obs[i,1],theta_obs[i,2])
      }
   }

   #Step 8: Estimate overall expected misclassifications
   inacc <- data.frame(inacc_type=inacc[,1],value=inacc[,2])
   Data[s] <- mean(inacc$value)

   #Step 9: Estimate the expected proportion of false positives
   #            and false negatives
```

```
    #9a. False positives
    falsepos <- inacc %>% filter(inacc_type==0)
    Data[s+nsimul] <- mean(falsepos$value)*(length(falsepos$value)/n)
    #9b. False negatives
    falseneg <- inacc %>% filter(inacc_type==1)
    Data[s+nsimul*2] <- mean(falseneg$value)*(length(falseneg$value)/n)

    #Remove objects from R workspace
    rm(falsepos,falseneg,inacc,resps,theta_obs,theta_sim)
}
```

Appendix 2. R Code Used to Conduct Simulations With two Cut Scores

This appendix shows the R code I used to simulate each condition with two cut scores. Once again, after step 3a of the code I manually paired item discriminations with item difficulty values in such a way as to maximize test information around the cut scores. Once done, the rest of code can be run.

```
#Simulate expected misclassifications with IRT 1PL or 2PL models
#Two cut scores

#Load necessary R packages
library(tidyverse)
library(irtoys)
library(Runuran)

#Set constants
nsimul <- 1000 #Number of replications of steps 4 to 9 according to #step 10
n <- 2000    #Number of test takers according to step 4
k <- 100    #Number of items

#Steps 1 and 2: Set the number and value of cut scores
cutscore1 <- -1.5
cutscore2 <- 0
#Step 3: Set the item parameter values
#3a: Sampling or setting item parameters
#Sampling item difficulty values
b_cs1 <- urnorm(50, cutscore1, 0.5, cutscore1-0.75, cutscore1+0.75)
b_cs2 <- urnorm(50, cutscore2, 0.5, cutscore2-0.75, cutscore2+0.75)
b <- c(b_cs1,b_cs2)
#Fixing guessing parameter values at zero
c <- rep(0,k)
#When using 1PL model, set discrimination with these two lines
a1pl <- 0.25   #Item discrimination value
```

```
a <- rep(a1pl,k)
#When using 2PL model, sample discrimination with these five lines
meanlog_cs1 <- -2   #Mean of sampled lognormal distribution
meanlog_cs2 <- -2   #Mean of sampled lognormal distribution
a_cs1 <- urlnorm(50, meanlog_cs1, 1, 0, 3)
a_cs2 <- urlnorm(50, meanlog_cs2, 1, 0, 3)
a <- c(a_cs1,a_cs2)
#3b: Creating table of item parameter values
#(column 1=a, column 2=b, column 3=c)
params <- cbind(a,b,c)
params <- data.matrix(params)

#Loop to perform 1000 replications
Data <- rep(NA,nsimul*7)
for (s in 1:nsimul) {
 #Step 4: Draw a sample of 2000 test takers from a
 #    standard normal distribution (~N(m=0,sd=1))
 theta_sim <- rnorm(n,0,1)
 #Step 5: Simulate 100 responses to dichotomously scored test items
 resps <- sim(params,theta_sim)
 #Step 6: Estimate test takers' maximum likelihood ability
 #    and standard error of measurement
 theta_obs <- mlebme(resps,params,method="ML")

 #Step 7: Estimate individual probability of misclassification
 inacc <- matrix(NA,n,3)
 for (i in 1:length(theta_sim)) {
  if (theta_obs[i,1]<cutscore1) {
   #If ability < cutscore1, label at performance level 1
   inacc[i,1] <- 1
   inacc[i,2] <- pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
   inacc[i,3] <- pnorm(Inf,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2])
   }
  else if (theta_obs[i,1]>=cutscore2) {
   #If ability >= cutscore2, label at performance level 3
   inacc[i,1] <- 3
   inacc[i,2] <- pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
   inacc[i,3] <- pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
  }
  else {
   #If cutscore1 <= ability < cutscore2, label at performance level 2
   inacc[i,1] <- 2
   inacc[i,2] <- pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
   inacc[i,3] <- pnorm(Inf,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2])
  }
 }

 #Step 8: Estimate overall expected misclassifications
 inacc <- data.frame(perf_level=inacc[,1],lower=inacc[,2],upper=inacc[,3])
 Data[s] <- mean(c(inacc$lower,inacc$upper))
```

```
 #Step 9: Estimate expected misclassifications
 #    at each performance level
 #9a. Misclassifications at performance level 1 (PL 1)
 pl1 <- inacc %>% filter(perf_level==1)
 #p(2|1) = False PL 2
 Data[s+nsimul] <- mean(pl1$lower)*((0.5*length(pl1$lower))/n)
 #p(3|1) = False PL 3
 Data[s+nsimul*2] <- mean(pl1$upper)*((0.5*length(pl1$upper))/n)

 #9b. Misclassifications at performance level 2 (PL 2)
 pl2 <- inacc %>% filter(perf_level==2)
 #p(1|2) = False PL 1
 Data[s+nsimul*3] <- mean(pl2$lower)*((0.5*length(pl2$lower))/n)
 #p(3|2) = False PL 3
 Data[s+nsimul*4] <- mean(pl2$upper)*((0.5*length(pl2$upper))/n)

 #9c. Misclassifications at performance level 3 (PL 3)
 pl3 <- inacc %>% filter(perf_level==3)
 #p(1|3) = False PL 1
 Data[s+nsimul*5] <- mean(pl3$lower)*((0.5*length(pl3$lower))/n)
 #p(2|3) = False PL 2
 Data[s+nsimul*6] <- mean(pl3$upper)*((0.5*length(pl3$upper))/n)

 #Remove objects from R workspace
 rm(pl1,pl2,pl3,inacc,resps,theta_obs,theta_sim)
}
```

EVALUAR

2021, Vol. 21, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

Invarianza factorial del Reactive/Proactive Aggression Questionnaire (RPQ) en adolescentes limeños institucionalizados y no institucionalizados

Factorial Invariance of the Reactive/Proactive Aggression Questionnaire in Institutionalized and Non-institutionalized Lima Adolescents

Rubén Gabriel Castañeda-Bernal [1] , Jossué David Correa-Rojas * [2] , Eli Leonardo Malvaceda-Espinoza [3]

1 - Universidad Nacional Mayor de San Marcos.
2 - Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
3 - Universidad San Ignacio de Loyola, Perú.

Recibido: 14/06/2021 **Revisado:** 11/08/2021 **Aceptado:** 16/09/2021

Resumen

La agresión es una problemática de importancia para los adolescentes, por ello, la comprensión de este fenómeno es crucial. El propósito de este estudio es analizar la validez estructural y la invarianza del Reactive/Proactive Aggression Questionnaire (RPQ) en adolescentes limeños institucionalizados y no institucionalizados. Se seleccionaron 344 adolescentes hombres, entre 15 y 17 años (M_{edad} = 16.055, DE_{edad} = .674), el 51.16% se encontraban institucionalizados en un centro juvenil de diagnóstico y rehabilitación social debido a conflictos con la ley penal. Los resultados muestran que el modelo bidimensional del RPQ presenta índices de ajuste relativamente aceptables (SB-χ^2 = $461.463_{(229)}$, CFI = .914, RMSEA = .054 [.047-.062]). Además, en ambos grupos, se estableció la invarianza configuracional, métrica, escalar y estricta. Se reportan coeficientes omega adecuados para la *agresión reactiva* (ω = .797) y *agresión proactiva* (ω = .837). Se concluye que el RPQ es una medida bidimensional, parsimoniosa e interpretable que mide la agresión reactiva y proactiva en los adolescentes ya mencionados.

Palabras clave: *validez, fiabilidad, agresión proactiva, agresión reactiva, violencia*

Abstract

Aggression is an important problem of adolescents, the understanding of this phenomenon is crucial. The purpose of this study is to analyze the structural validity and invariance of the Reactive/Proactive Aggression Questionnaire (RPQ) in institutionalized and non-institutionalized adolescents from Lima. 344 male adolescents were selected, among 15 and 17 years old (M_{age} = 16.055, SD_{age} = .674), 51.16% were institutionalized in a youth center for diagnosis and social rehabilitation due to conflicts with criminal law. The results show that the two-dimensional model of the RPQ presents relatively acceptable fit indices (SB-χ^2 = $461.463_{(229)}$, CFI = .914, RMSEA = .054 [.047-.062]). Additionally, the configurational, metric, scalar and strict invariance was established in both groups. Adequate omega coefficients are reported for *reactive aggression* (ω = .797) and *proactive aggression* (ω = .837). It is concluded that the RPQ is a two-dimensional, parsimonious, and interpretable measure that determines reactive and proactive aggression in the aforementioned adolescents.

Keywords: *validity, reliability, proactive aggression, reactive aggression, violence*

***Correspondencia a:** Jossué Correa Rojas. Prolongación Primavera 2390, Monterrico, Santiago de Surco - Lima, Perú. E-mail: jossue.correa@upc.pe

Cómo citar este artículo: Castañeda-Bernal, R. G., Correa-Rojas, J. D., & Malvaceda-Espinoza, E. L. (2021). Invarianza factorial del Reactive/Proactive Aggression Questionnaire (RPQ) en adolescentes limeños institucionalizados y no institucionalizados. *Revista Evaluar, 21*(3), 35-48. Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Participaron en la edición de este artículo: Florencia Ruiz, Alicia Molinari, Mónica Serppe, Andrea Suárez, Juan Balverdi, Eugenia Barrionuevo, Ricardo Hernández.

Introducción

Durante el primer trimestre de 2019 se registraron más de 4500 casos de violencia escolar en Lima Metropolitana. Estos casos incluyen violencia física, psicológica y sexual. De estos, el 50.9% se dio entre los escolares de nivel secundario (Ministerio de Educación, 2019). Por lo tanto, resulta pertinente entender a la violencia como una expresión extrema de la agresión (Allen & Anderson, 2017).

La agresión es un comportamiento que busca dañar a otra persona que se encuentra motivada a evitarlo. Por lo tanto, es intencional, dado que se constituye como un acto para causar daño con consciencia de ello. La agresión implica diferentes conductas que, a pesar de parecer similares, poseen diferentes motivaciones (Allen & Anderson, 2017). Para objeto de la presente, se utilizará la distinción entre agresión *proactiva* y *reactiva* (Dodge, 1991).

La *agresión reactiva o impulsiva* es una conducta que se expresa como reacción a una provocación o amenaza percibida (real o imaginada) en diferentes situaciones (Andreu, Ramírez, & Raine, 2006). Es entendida por quien la ejecuta con propósito defensivo y está relacionada con la hostilidad (Roncero, Andreu, & Peña, 2016). A su vez, de acuerdo con Penado, Andreu y Peña (2014), se asocia con la impulsividad.

Por su parte, la *agresión proactiva* es el conjunto de conductas intencionadas y motivadas que tienen como fin causar daño a otra persona. Ello evidencia una evaluación positiva de la agresión (Ramírez & Andreu, 2006). Asimismo, quienes la ejercen pueden deshumanizar a sus víctimas (Penado et al., 2014). Por lo general, estas conductas agresivas se encuentran asociadas al trastorno disocial (Lobbestael, Cima, & Lemmens, 2015), así como a la conducta antisocial (Penado et al., 2014).

Diversos estudios realizados en España con estudiantes adolescentes no institucionalizados señalan que la conducta agresiva puede afectar el desarrollo de la autoestima y el autoconcepto (Torregrosa, Inglés, & García-Fernández, 2011). Por otro lado, Garaigordobil, Durá y Pérez (2005) encontraron que la conducta agresiva favorece la conducta antisocial en adolescentes. Asimismo, Torregrosa, Inglés, García-Fernández, Valle y Núñez (2012) señalaron que la conducta agresiva afecta significativamente las metas académicas de los escolares de nivel secundario.

De acuerdo con lo señalado, y considerando que la adolescencia es una etapa del ciclo vital relativamente sensible a los cambios (Andolfi & Mascellani, 2012), es de suma importancia conocer cómo los adolescentes expresan sus conductas agresivas (Linares, 2015). De esta manera, Cerezo-Ramírez y Méndez-Mateo (2009) en un estudio realizado con adolescentes españoles no institucionalizados, señalan que la mayoría de ellos no presentan más que los problemas propios de esta etapa; sin embargo, un grupo significativo de adolescentes suele iniciar problemas de conducta que pueden tornarse crónicos. Esto último fue reportado por Vega-Cauich y Zumárraga-García (2019), en un estudio con adolescentes institucionalizados debido a conflictos con la ley penal, en México.

La literatura especializada coincide en que los adolescentes de sexo masculino son más propensos a manifestar estas conductas agresivas (Bernardes de Moraes, 2013). Esto fue corroborado por Torregrosa et al. (2010) quienes lo identificaron en estudiantes adolescentes españoles, chinos y mexicanos no institucionalizados. En Perú, la predominancia de las conductas agresivas en estudiantes adolescentes varones es elevada (Rojas-Zegarra et al., 2020; Romaní, Gutiérrez, & Lama, 2011). Esto revela la importancia del estudio de la conducta agresiva en los adolescentes,

ya que su establecimiento en edades tempranas suele estructurar pautas de interacción inadecuadas que pueden generar consecuencias negativas (De la Cruz-Gil, 2008).

Desde este marco, el estudio de la agresión se ha basado en el modelo teórico de Dodge (1991), ya que es aceptado por la comunidad académica para explicar la agresión en adolescentes. A partir de sus postulados, Raine et al. (2006) diseñaron el Reactive/Proactive Aggression Questionnaire (RPQ), sobre una muestra de 334 adolescentes estadounidenses. El análisis factorial confirmatorio (AFC) corroboró que la RPQ es una medida bidimensional (*proactive-reactive model*) capaz de discriminar la agresión reactiva y proactiva en muestras de adolescentes. Las cargas factoriales para *reactive factor* fluctuaron entre .48 y .65, mientras que para *proactive factor* las mismas variaron entre .52 y .64, con índices de ajuste adecuados ($\chi^2_{(gl)} = 334_{(229)}$; CFI = .91; RMSEA = .04). Asimismo, determinaron la fiabilidad de la medida mediante el coeficiente alfa el cual oscila entre .84 (*reactive factor*) y .90 (*proactive factor*).

Posteriormente, este instrumento fue adaptado al español por Andreu-Rodríguez, Peña-Fernández y Ramírez (2009) sobre una muestra de 732 adolescentes españoles. La validación se realizó a través de una AFC con el método de mínimos cuadrados no ponderados, y se demostró que el modelo bifactorial presentaba índices de ajuste adecuados, respecto a un modelo unidimensional (GFI = .98; NFI = .96; RMR = .02). Las cargas factoriales para el factor de agresión reactiva se encontraron alrededor de .49 y .70, mientras que para el factor de agresión proactiva estas fluctuaron entre .50 y .72. El hallazgo más interesante fue que los varones presentaron mayores niveles de agresión proactiva que las mujeres.

El RPQ también ha sido analizado en otros contextos, y ha mostrado una adecuada prestancia para medir el constructo en adolescentes portugueses institucionalizados en centros de rehabilitación social debido a conflictos con la ley penal (Pechorro, Ray, Raine, Marocco, & Gonçalves, 2015). Así también, Cenkseven-Önder, Avcı y Çolakkadıoğlu (2016) adaptaron el instrumento en adolescentes turcos, constatando la validez y fiabilidad del instrumento. Del mismo modo, Penado et al. (2014) analizaron las propiedades psicométricas del RPQ y demostraron la consistencia de sus medidas en escolares españoles.

En Perú, Abanto-Chomba (2018) realizó la validación del RPQ en estudiantes adolescentes no institucionalizados. Los resultados indican que el RPQ presenta un modelo bidimensional con un ajuste general adecuado con índices RMSEA = .05, SRMR = .05, CFI = .91, TLI = .90, satisfactorios. Con respecto a la fiabilidad, se reportan coeficientes omega cuyos valores fluctúan entre .73 (*reactiva*) y .77 (*proactiva*). Del mismo modo, Florián-Guarniz (2018) adaptó el RPQ en escolares peruanos de la ciudad de Huaraz y reportó resultados similares en cuanto a la bidimensionalidad del instrumento. La estimación de la fiabilidad sugiere que el instrumento original presenta una adecuada consistencia en la medición de la agresividad reactiva (ω = .66) y proactiva (ω = .70).

Un estudio realizado por Rojas-Zegarra et al. (2020) llevado a cabo sobre una muestra de 2830 adolescentes peruanos no institucionalizados de la ciudad de Arequipa da cuenta de la validez y fiabilidad del RPQ. Los resultados evidencian la validez estructural del modelo bidimensional, el cual se obtuvo por el método de mínimos cuadrados ponderados diagonalmente (DWLS) que arrojó cargas factoriales que fluctúan entre .52 y .77 (*reactiva*) y .49 y .81 (*proactiva*). En dicha investigación, los índices de ajuste fueron mayormente adecuados para validar el instrumento original (CFI y TLI mayores de .90 y RMSEA y SRMR menores de .08).

Si bien existen investigaciones que han explorado las evidencias de validez y fiabilidad del RPQ en adolescentes peruanos (Rojas-Zegarra et al., 2020), no se han realizado validaciones específicamente en adolescentes institucionalizados debido a conflictos con la ley penal. Lo anterior es de suma importancia, ya que según el Observatorio Nacional de Política Criminal (2017), dichos adolescentes institucionalizados presentan problemas de ansiedad, depresión y adicción a sustancias. Asimismo, sus entornos familiares son desorganizados, han crecido en abandono funcional y en situaciones de riesgo. Debido a la problemática que todo ello implica es necesario realizar estudios que permitan caracterizar, comparar, modelar y predecir tales conductas.

En tal sentido el objetivo del presente estudio es analizar la validez estructural e invarianza del Reactive/Proactive Questionnaire (RPQ) en adolescentes limeños no institucionalizados e institucionalizados debido a conflictos con la ley penal. Al respecto, el establecimiento de la invarianza factorial resulta necesario pues es una propiedad de los instrumentos de medida que permite realizar comparaciones entre grupos que presentan una condición que podría generar un sesgo en la medición (Byrne, 2008).

Método

Participantes

Se trata de un estudio instrumental (Ato, López-García, & Benavente, 2013), para el cual se seleccionaron intencionalmente 344 adolescentes varones de Lima Metropolitana, cuyas edades están comprendidas entre los 15 y 17 años (M_{edad} = 16.06, DE_{edad} = .67), de los cuales el 51.16% se encontraban internados en un centro juvenil de diagnóstico y rehabilitación social debido a que cometieron alguna infracción a la ley penal (p. ej. homicidio, robo, hurto y venta ilícita de drogas; M = 16.14, DE = .77) y 48.84% provienen de una institución educativa de nivel secundario (M = 15.97, DE = .55). Estos últimos fueron elegidos debido a su accesibilidad y nivel sociocultural, el mismo que resulta similar al del grupo de adolescentes institucionalizados, como es el caso de provenir de zonas urbanas de Lima y de un estrato social bajo. Cabe mencionar que los jóvenes institucionalizados se encontraban culminando el nivel primario y cursos de los dos primeros años del nivel secundario y los jóvenes no institucionalizados se encontraban cursando niveles de estudios similares.

Respecto a la idoneidad del tamaño de muestra, Herrero (2010) señala que no existe un consenso respecto al tamaño de muestra para los modelos SEM; sin embargo, indica que la fiabilidad del modelo depende de su complejidad y del número de sujetos con que cuenta el investigador para contrastarlo, pues en ello radica la complejidad del modelo y de si se han realizado modificaciones *post-hoc* en el mismo. Asimismo, Lloret-Segura, Ferreres-Traver, Hernández-Baeza y Tomás-Marco (2014) sostienen que solo en los casos en los cuales la medida haya presentado comunalidades en torno al .30 y con tres ítems por factor, serán necesarias muestras mayores a 400 participantes.

Instrumento

Cuestionario de Agresión Reactivo-Proactivo (CAR-P). Para medir la agresión reactiva y proactiva se utilizó el CARP, diseñado por Raine et al. (2006), el cual consta de 23 ítems distribuidos en dos dimensiones: *agresión reactiva* (AR) y *agresión proactiva* (AP), con opciones de respuesta

tipo Likert: (1) *nunca* (2) *a veces* y (3) *a menudo*. El instrumento ha sido validado en otros países como España y Turquía. En la presente investigación, se empleó la adaptación en Perú realizada por Rojas-Zegarra et al. (2020), la cual fue aplicada a 2830 estudiantes de nivel secundario de 13 a 19 años. La evidencia de validez se determinó mediante un análisis factorial confirmatorio con el método de mínimos cuadrados diagonalmente ponderados (DWLS), y se encontraron cargas factoriales estandarizadas para ambos factores que oscilan entre .31 y .77. Los índices de ajuste resultaron satisfactorios, con valores RMSEA = .07, SRMR = .07, CFI = .96, TLI = .96. Se evaluó la confiabilidad mediante el método de consistencia interna, reportando coeficientes omegas por encima de .70 para ambos factores.

Procedimiento

Inicialmente se solicitaron los permisos respectivos a las instituciones en las cuales se realizó la investigación. Una vez recibida la autorización para la ejecución del estudio, se llevaron a cabo las evaluaciones en las instalaciones de una institución educativa del distrito del Rimac. Simultáneamente se iniciaron las evaluaciones en el Centro Juvenil de Diagnóstico y Rehabilitación ubicado en el distrito de San Miguel. Ambas instituciones se encuentran en Lima Metropolitana. Las mediciones se llevaron a cabo entre los meses de abril y noviembre del 2019. Previo a la administración del RPQ, los participantes firmaron el consentimiento informado. En este documento se dio a conocer el carácter voluntario del estudio, la libertad de su participación, la ausencia de daño físico y psicológico y la confidencialidad de la información recabada. En tal sentido, se siguieron las recomendaciones de la American Educational Research Association (AERA), American Psychological Association (APA) y el National Council on Measurement in Education (NCME; AERA, APA, & NCME 2014). Finalmente, se informó a los participantes el propósito de la evaluación y se indicó que podían solicitar sus resultados de forma individual, con total confidencialidad.

Análisis de datos

El análisis estadístico se realizó a través de una serie de etapas. En la primera se analizaron las medidas descriptivas de los ítems y sus características distribucionales, siendo evaluada la normalidad a través de los coeficientes de asimetría y curtosis. Se considera a los valores dentro del rango de ± 1.5 como indicadores de normalidad univariada (Pérez & Medrano, 2010). Luego, para identificar las evidencias de validez estructural, se realizó un análisis factorial confirmatorio (AFC) con el método *weighted least square mean and variance adjusted* (WLSMV), debido a la naturaleza categórica de las variables de estudio (Domínguez-Lara, 2014; Verdam, Oort, & Sprangers, 2016), además de ser un estimador más confiable en muestras pequeñas (Li, 2014), se consideraron pesos factoriales aceptables a partir de .40 (Williams, Onsman, & Brown, 2010). Asimismo, se evaluaron los índices de ajuste del modelo, entre ellos: la razón chi cuadrado sobre los grados de libertad (χ^2/gl) con valores esperados menores a 3, *root mean square error of approximation* (RMSEA) y *standardized root mean square residual* (SRMR). En ambos casos se esperan valores por debajo de .08 sugeridos por Bentler y Bonnet (1980).

Se incluyó el *comparative fit index* (CFI) de Jöreskog y Sörbom (1986) y el índice de Tuc-

ker-Lewis (TLI) ambos con valores esperados por encima de .90 (Kline, 2015; Hair, Andreson, Tatham, & Black, 1999). Adicionalmente, a partir de las cargas factoriales, se calculó la varianza promedio extraída (*average variance extracted* [AVE]) considerando valores alrededor del .50 como satisfactorios. Con ello se verificaron las evidencias de validez interna convergente (Fornell & Larcker, 1981). Para establecer la invarianza de la medida en diferentes grupos, se utilizó el criterio de Wu y Estabrook (2016), quienes consideran un nivel estándar y una parametrización *theta*. Asimismo, los autores realizan una variación a la invarianza configuracional, fijando los *thresholds* (umbrales). Para la invarianza métrica se fijaron los umbrales y las cargas factoriales. En el caso de la invarianza escalar se fijaron los umbrales, las cargas factoriales y los interceptos.

Luego, para determinar la invarianza estricta se restringieron los umbrales, cargas factoriales, interceptos y residuos. La invarianza de la medida se evaluó a través de los cambios menores a .01 en los índices CFI (Byrne, 2008). También se consideraron cambios en el RMSEA (ΔRMSEA) ≤ .01, y en el SRMR (ΔSRMR) ≤ .03 asumiendo estos criterios como medidas adecuadas para aceptar la invarianza (Chen, 2007). La confiabilidad se evaluó en su consistencia interna con el coeficiente omega categórico ($\omega_{categoórico}$) obtenido a través del BCA *bootstrap* junto a sus intervalos de confianza (IC) al 95% (Ventura-León, 2018a).

Para establecer la validez por grupos contrastables se compararon los puntajes de la AR y AP mediante el estadístico *t* de Student, se estableció para su interpretación un error estimado el .05 y un nivel de confianza del 95%. Asimismo, se calcularon los tamaños de efecto d de Cohen para comparación de muestras independientes, cuyos valores pueden ser interpretados como pequeño ($d > .20$), mediano ($d > .50$) o grande ($d > .80$; Cohen, como se citó en Ventura-León, 2018b).

Finalmente, la confiabilidad compuesta se calculó a partir de la sumatoria cuadrática de las cargas factoriales entre los errores de medida, la cual considera los cambios en las cargas factoriales producto de la inclusión de errores correlacionados entre los ítems (Hair et al., 1999).

Para los análisis se utilizaron el programa IBM SPSS, versión 25 y el RStudio versión 3.3.2 (RStudio Team, 2015), empleándose el paquete Lavaan (Rosseel et al., 2021).

Resultados

Análisis descriptivo

En la Tabla 1, se presenta el análisis descriptivo de los ítems que componen el CAR-P. Las medidas reportadas dan cuenta de la media (M), desviación estándar (DE), coeficiente de asimetría (g_1) y curtosis (g_2). Estos valores se calcularon a partir de las puntuaciones obtenidas en cada uno de los ítems que conforman el instrumento. Se evidenció que los ítems 1 (M = 1.98) y 19 (M = 1.90) presentan las medias aritméticas más altas, mientras que las medias más bajas están presentes en los ítems 18 (M = 1.21) y 21 (M = 1.21). En cuanto a la variabilidad, se aprecia que los ítems 14 (DE = .69) y 19 (DE = .66) son los que presentan mayor dispersión. La asimetría y curtosis arrojan valores por encima de ± 1.50, lo que indica que la distribución de los ítems no se aproxima a una distribución univariante normal (Pérez & Medrano, 2010).

Evidencias de validez basada en la estructura interna

En la Tabla 2, respecto a la estructura inter-

Tabla 1
Estadísticos descriptivos.

	Mín	Máx	M	DE	g_1	g_2
P1	1.000	3.000	1.983	.523	-0.022	0.685
P2	1.000	3.000	1.459	.605	0.953	-0.099
P3	1.000	3.000	1.904	.591	0.025	-0.182
P4	1.000	3.000	1.721	.655	0.362	-0.741
P5	1.000	3.000	1.843	.651	0.168	-0.682
P6	1.000	3.000	1.453	.633	1.080	0.069
P7	1.000	3.000	1.657	.580	0.222	-0.674
P8	1.000	3.000	1.529	.634	0.789	-0.397
P9	1.000	3.000	1.381	.594	1.309	0.680
P10	1.000	3.000	1.308	.570	1.697	1.861
P11	1.000	3.000	1.593	.613	0.518	-0.623
P12	1.000	3.000	1.323	.564	1.563	1.467
P13	1.000	3.000	1.837	.622	0.126	-0.508
P14	1.000	3.000	1.808	.690	0.272	-0.896
P15	1.000	3.000	1.308	.570	1.697	1.861
P16	1.000	3.000	1.308	.570	1.697	1.861
P17	1.000	3.000	1.311	.523	1.421	1.078
P18	1.000	3.000	1.206	.472	2.254	4.413
P19	1.000	3.000	1.901	.658	0.107	-0.698
P20	1.000	3.000	1.323	.527	1.350	0.863
P21	1.000	3.000	1.209	.498	2.369	4.798
P22	1.000	3.000	1.297	.550	1.705	1.963
P23	1.000	3.000	1.186	.439	2.319	4.811

na del instrumento, se aprecia que la mayoría de los reactivos alcanzan cargas factoriales satisfactorias (> .40) tal como sugiere la literatura (Williams et al., 2010), con excepción del ítem 18, el cual obtiene una carga factorial igual a .33 y que se ubica en la dimensión *agresión proactiva* (AP). Los valores de los residuos fluctúan entre .41 y .81 para la dimensión *agresión reactiva* (AR), mientras que estos oscilan entre .33 y .89 para la dimensión AP. El promedio de la suma de las cargas factoriales al cuadrado (AVE) para el caso de AR es igual a .37, para AP el AVE es igual a .41, cuyo valor se encuentra ligeramente por debajo de lo sugerido (AVE > .50), por lo que no se puede precisar que la estructura bidimensional original presente validez interna convergente (Fornell & Larcker, 1981). A partir de las cargas factoriales se calculó el coeficiente de fiabilidad compuesta (FC) para AR (FC = .86) y AP (FC = .75), ambos casos corresponden a valores adecuados (Hair et al., 1999). En cuanto a los índices de ajuste del modelo robusto, se reporta un SB-χ^2 =

Tabla 2
Estructura factorial del RPQ.

Ítems	AR	AP	λ^2	*e*
P1	.57		.32	.68
P3	.66		.43	.57
P5	.43		.19	.81
P7	.50		.25	.75
P8	.65		.42	.58
P11	.77		.59	.42
P13	.48		.23	.77
P14	.52		.27	.73
P16	.77		.59	.41
P19	.52		.27	.73
P22	.75		.56	.44
P2		.66	.43	.57
P4		.58	.33	.67
P6		.62	.38	.62
P9		.62	.39	.61
P10		.77	.59	.41
P12		.68	.46	.54
P15		.68	.47	.53
P17		.61	.37	.63
P18		.33	.11	.89
P20		.61	.37	.63
P21		.82	.67	.33
P23		.59	.35	.65
	F1	F2		
F1	-		-	-
F2	.79	-	-	-
AVE	.37	.41	-	-
FC	.86	.75	-	-

461.46 (gl = 229; p < .01), con índices de ajuste comparativo aceptables (CFI = .91, TLI = .91) y el RMSEA es igual a .05 [.05-.06], ambos correspondientes a un ajuste adecuado (Ruiz, Pardo, & San Martín, 2010), al igual que el SRMR = .08 (Hair et al., 1999).

Invarianza de la medida según condición de residencia

Adicionalmente, se analizó la invarianza del RPQ, cuyos resultados muestran que los ΔCFI, ΔRMSEA y ΔSRMR se encuentran dentro de los

parámetros sugeridos (Byrne, 2008; Chen, 2007). Estos hallazgos permiten establecer la invarianza configuracional, métrica, escalar y estricta de la medida de la agresión reactiva y proactiva en adolescentes institucionalizados y no institucionalizados (ver Tabla 3).

Tabla 3
Invarianza del RPQ según condición de residencia.

Invarianza	$\chi^2_{(gl)}$	CFI	ΔCFI	RMSEA [IC 90%]	ΔRMSEA	SRMR	ΔSRMR
Configuracional*	$679.76_{(458)}$	.93	-	.05 [.05 - .06]	-	.11	-
Métrica	$694.12_{(479)}$	.93	.00	.05 [.04 - .06]	.00	.11	.00
Escalar	$760.72_{(500)}$	.91	.02	.05 [.05 - .06]	.00	.12	.01
Estricta	$785.84_{(523)}$	.91	.00	.05 [.05 - .06]	.00	.12	.00

Nota. * con *threshold* fijados según el método de Wu y Estabrook (2016).

Evidencias de fiabilidad

La fiabilidad del RPQ se evaluó por el método de consistencia interna. Se reporta el coeficiente omega categórico ($\omega_{categórico}$) con sus respectivos intervalos de confianza, obteniéndose ω = .80 [IC = .76, - .83] para la *agresión reactiva* (AR) y ω = .84 [IC = .80, - .87] para la *agresión proactiva* (AP) y también se calculó el coeficiente omega para los subgrupos. Los adolescentes institucionalizados obtuvieron un coeficiente igual a .83 para AR y .86 para AP, mientras que los adolescentes no institucionalizados alcanzaron un omega igual .74 para AR y .55 para AP.

Evidencias de validez por grupos contrastables

Se realizó un análisis de validez por grupos contrastados. Para ello se comparó la agresión reactiva y proactiva en adolescentes institucionalizados y no institucionalizados. Los resultados se exponen en la Tabla 4, se observa que no se identificaron diferencias estadísticamente significativas entre ambos grupos, con un efecto pequeño (*d* = .17). Sin embargo, sí se encontraron diferencias estadísticamente significativas para los dos grupos en lo referente a la *agresión proactiva*, con un tamaño de efecto mediano (*d* = .67).

Discusión

El presente estudio tuvo como objetivo analizar la validez estructural y la invarianza del RPQ en adolescentes no institucionalizados e institucionalizados. Los hallazgos demuestran que la RPQ es una medida bidimensional, coherente,

Tabla 4
Comparación de la agresión reactiva y proactiva, según condición.

Variable	M(DE) $_{G1}$	M(DE) $_{G2}$	*t*	*p*	*d*
Agresión reactiva	18.98 (4.30)	18.32 (3.29)	1.61	> .05	.17
Agresión proactiva	17.39 (4.73)	14.93 (2.49)	6.09	< .01	.65

parsimoniosa e interpretable capaz de discriminar la agresión reactiva y proactiva en adolescentes, lo cual es coherente con el modelo original de Raine et al. (2006).

No obstante, una diferencia del presente estudio con respecto a los otros (Raine et al., 2006; Rojas-Zegarra et al., 2020) es la muestra con la que se han explorado las evidencias de validez y fiabilidad del RPQ. Un grupo estuvo conformado por adolescentes institucionalizados, que se encuentran internados en un centro juvenil de diagnóstico y rehabilitación, por haber cometido infracciones a la ley peruana y otro grupo por adolescentes que se desenvuelven en condiciones regulares en un centro educativo (adolescentes no institucionalizados).

En cuanto a las evidencias de validez basadas en la estructura interna, no solo se corrobora la estructura teórica del instrumento, sino que se obtuvieron cargas factoriales sustancialmente superiores en ambos factores en comparación con el estudio de Raine et al. (2006) llevado a cabo en adolescentes españoles y con otro estudio desarrollado con una muestra de adolescentes peruanos (Rojas-Zegarra et al., 2020). Esto puede deberse al método de extracción utilizado. Mientras que en los estudios mencionados se emplearon métodos como máxima verosimilitud o mínimos cuadrados ponderados diagonalmente, en esta investigación el método utilizado fue WLSMV debido a la naturaleza categórica de las variables (Domínguez-Lara, 2014; Verdam et al., 2016), además de ser un estimador más confiable en muestras pequeñas (Li, 2014).

Asimismo, los resultados permitieron establecer la invarianza del RPQ en adolescentes institucionalizados y no institucionalizados, lo que demuestra la equivalencia de su estructura interna y de sus puntuaciones en ambos grupos (Byrne, 2008; Wu & Estabrook, 2016). Con ello se corrobora que la medida no presenta sesgos de medición con respecto a esta condición de los adolescentes. Sin embargo, este resultado no es comparable con otros estudios puesto que no se han identificado investigaciones que hayan evaluado estos parámetros.

En este sentido, al establecerse la invarianza configuracional, se halló que la RPQ es una medida que sostiene la misma organización del constructo para los grupos. Asimismo, al demostrarse la invarianza métrica, se entiende que cada reactivo compone la medición del constructo en grado similar (Putnick & Bornstein, 2016). Al establecerse la invarianza escalar, es posible realizar comparaciones que permitan diferenciar el predominio de agresión en uno y otro grupo con la seguridad de que las diferencias encontradas corresponden a la condición de los adolescentes y no a sesgos en el instrumento (Lee, 2018). Al demostrar la invarianza estricta (residual), se demuestra que la varianza específica y error son similares en ambos grupos (Elosua, 2005).

Estos resultados tienen implicancias teóricas y prácticas para la investigación. A nivel teórico, al corroborar la estructura bidimensional de la RPQ, se provee de evidencia empírica al modelo teórico subyacente sobre el cual fue diseñado originalmente el instrumento (Nunnally, 2013). A nivel práctico, estos hallazgos justifican el uso de una medida psicológica para identificar la agresión reactiva y proactiva de forma diferenciada. Además, puede utilizarse en diferentes contextos, tanto clínicos como educativos (Medrano & Pérez, 2019). Finalmente, al identificar que se trata de una medida que cuenta con evidencias de validez y fiabilidad, su uso en la investigación garantiza mediciones libres de sesgos producidos por errores sistemáticos (Bisquerra, 2004)

En lo que respecta a la fiabilidad del RPQ, se obtuvo una consistencia interna similar a la reportada en otras investigaciones (Raine et al., 2006; Rojas-Zegarra et al., 2020). Los hallazgos

confirman la estabilidad de las mediciones del RPQ tanto en los adolescentes institucionalizados como en los adolescentes no institucionalizados. Sin embargo, se puede apreciar que la RPQ presentó medidas más consistentes en el grupo de adolescentes institucionalizados, posiblemente porque en ellos los efectos de la deseabilidad dada su condición se hicieron menos relevantes. Esta hipótesis se confirma al comparar la agresión reactiva y proactiva en ambos grupos, pues solo se encontraron diferencias significativas en la agresión proactiva a favor del grupo de adolescentes institucionalizados, quienes por su condición se espera que presenten este atributo (Allen & Anderson, 2017) pues está íntimamente relacionado con conductas antisociales (Penado et al., 2014) propias de la circunstancia que ha llevado a su internamiento.

Al haber constatado la validez estructural y la invarianza del RPQ, se viabiliza el desarrollo de futuros estudios que puedan explorar las diferencias entre adolescentes institucionalizados y no institucionalizados. Asimismo, permite establecer en qué medida la agresión reactiva y proactiva podrían constituir factores de riesgo para la reincidencia delictiva (Horcajo-Gil, Dujo-López, Andreu-Rodríguez, & Marín-Rullán, 2019).

Entre las principales limitaciones de la investigación hemos de mencionar que los resultados no pueden ser generalizables, siendo interpretables únicamente en relación a la particularidad de la muestra seleccionada debido al tipo de muestreo utilizado. También, por la naturaleza del estudio y por los permisos obtenidos, no fue posible establecer otros tipos de evidencia de validez. Asimismo, no se analizó la invarianza factorial del RPQ considerando otros aspectos sociodemográficos como la edad y el tiempo de internamiento.

En cuanto a las recomendaciones, es necesario que se incremente la cantidad de participantes considerando otras características sociodemográficas (otros estratos económicos y sociales) además de plantear un tipo de muestreo probabilístico. Es pertinente plantear la invarianza de medición a nivel regional debido a las marcadas diferencias de esta característica sociodemográfica en el Perú. Por último, se considera necesario, en futuras investigaciones, analizar las evidencias de validez basadas en la relación con otras variables y estimar las evidencias de fiabilidad mediante el método de estabilidad temporal. Adicionalmente, es necesario plantear un modelo explicativo para conocer las causas y consecuencias de la agresión proactiva en los adolescentes institucionalizados.

Referencias

Abanto-Chomba, A. L. (2018). *Evidencias de Validez del cuestionario de Agresión Reactiva y Proactiva (RPQ) en adolescentes de Huaraz* (Tesis de pregrado). Recuperado de https://repositorio.ucv.edu.pe

Allen, J. J., & Anderson, C. A. (2017). Aggression and violence: Definitions and distinctions. En P. Sturmey (Ed.), *The Wiley Handbook of Violence and Aggression* (pp. 1-14). doi: 10.1002/9781119057574.whbva001

Andolfi, M., & Mascellani, A. (2012). *Historias de la adolescencia.* Barcelona: Gedisa.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington D.C.: American Educational Research Association.

Andreu, J. M., Ramirez, J. M., & Raine, A. (2006). Un modelo dicotómico de la agresión: Valoración mediante dos auto-informes (CAMA y RPQ). *Psicopatología Clínica, Legal y Forense, 6*, 25-42. Recuperado de https://dialnet.unirioja.es/servlet/revista?codigo=5657

Andreu-Rodríguez, J. M., Peña-Fernández, M. E., & Ra-

mírez, J. M. (2009). Cuestionario de agresión reactiva y proactiva: Un instrumento de medida de la agresión en adolescentes. *Revista de Psicopatología y Psicología Clínica, 14*(1), 37-49. doi: 10.5944/rppc.vol.14.num.1.2009.4065

Ato, M., López-García, J. J., & Benavente, A. (2013). Un sistema de clasificación de los diseños de investigación en psicología. *Revista Anales de Psicología, 29*(3), 1038-1051. doi: 10.6018/analesps.29.3.178511

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606. doi: 10.1037/0033-2909.88.3.588

Bernardes de Moraes, T. (2013). ¿Por qué los hombres presentan un comportamiento más agresivo que las mujeres? Por una antropología evolutiva del comportamiento agresivo. *Nómadas. Critical Journal of Social and Juridical Sciences, 37*(1), 93-111. doi: 10.5209/rev_NOMA.2013.v37.n1.42561

Bisquerra, R. (2004). *Metodología de la investigación educativa*. Madrid: La Muralla.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882. Recuperado de http://www.psicothema.com

Cenkseven-Önder, F., Avcı, R., & Çolakkadıoğlu, O. (2016). Validity and reliability of the Reactive Proactive Aggression Questionnaire in Turkish adolescents. *Educational Research and Reviews, 11*(20), 1931-1943. doi: 10.5897/ERR2016.2937

Cerezo-Ramírez, F., & Méndez-Mateo, I. (2009). Adolescentes, agresividad y conductas de riesgo de salud: Análisis de variables relacionadas. *International Journal of Developmental and Educational Psychology, 1*(1), 217-225. Recuperado de https://www.redalyc.org/pdf/3498/349832320023.pdf

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504. doi: 10.1080/10705510701301834

De la Cruz-Gil, R. (2008). *Violencia Intrafamiliar. Enfoque Sistémico.* Ciudad de México: Trillas.

Dodge, K. A. (1991). The structure and function of reactive and proactive aggression. En D. J. Pepler & K. H. Rubin (Eds.), *The development and treatment of childhood aggression* (pp. 201-218). Hillsdale, NJ: Earlbaum.

Domínguez-Lara, S. A. (2014). ¿Matrices policóricas/tetracóricas o matrices Pearson? Un estudio metodológico. *Revista Argentina de Ciencias del Comportamiento, 6*(1), 39-48. Recuperado de https://revistas.unc.edu.ar/index.php/racc/index

Elosua, P. (2005). Evaluación progresiva de la invarianza factorial entre las versiones original y adaptada de una escala de autoconcepto. *Psicothema, 17*(2), 356-362. Recuperado de http://www.psicothema.com

Florián-Guarniz, D. V. (2018). *Evidencias de validez del Cuestionario de Agresión Reactiva y Proactiva en adolescentes del distrito de Contumazá, 2018* (Tesis de grado). Recuperado de https://repositorio.ucv.edu.pe

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. doi: 10.1177/002224378101800104

Garaigordobil, M., Durá, A., & Pérez, J. I. (2005). Síntomas psicopatológicos, problemas de conducta y autoconcepto-autoestima: Un estudio con adolescentes de 14 y 17 años. *Anuario de Psicología Clínica y de la Salud, 1*, 53-63. Recuperado de http://institucionales.us.es/apcs/php/index.php?option=com_content&view=article&id=66&Itemid=11

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1999). *Análisis multivariante* (2ª ed.). Madrid: Prentice Hall.

Herrero, J. (2010). El análisis factorial confirmatorio en el estudio de la estructura y estabilidad de los instrumentos de evaluación: Un ejemplo con el Cuestionario de Autoestima CA-14. *Psychosocial Intervention, 19*(3), 289-300. Recuperado de https://journals.copmadrid.org/pi

Horcajo-Gil, P. J., Dujo-López, V., Andreu-Rodríguez, J.

M., & Marín-Rullán, M. (2019). Valoración y gestión del riesgo de reincidencia delictiva en menores infractores: Una revisión de instrumentos. *Anuario de Psicología Jurídica, 29*(1), 41-53. doi: 10.5093/apj2018a15

Jöreskog, K. G., & Sörbom, D. (1986). *Lisrel VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods.* Mooresville: Scientific Software.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling* (4ª ed.). New York: Guilford Press.

Lee, S. T. H. (2018). Testing for measurement invariance: Does your measure mean the same thing for different participants? *Observer, 31*(8), 32-33. Recuperado de https://www.psychologicalscience.org/issue/sept-oct21

Li, C. H. (2014). The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables (Tesis de doctorado). Recuperado de https://msu.edu

Linares, J. L. (2015). Prácticas alienadoras parentales. *El "síndrome de alienación parental" reformulado.* Barcelona: Gedisa.

Lobbestael, J., Cima, M., & Lemmens, A. (2015). The relationship between personality disorder traits and reactive versus proactive motivation for aggression. *Psychiatry Research, 229*(1-2), 155-160. doi: 10.1016/j.psychres.2015.07.052

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los items: Una guía práctica, revisada y actualizada. *Anales de Psicología, 30*(3), 1151-1169. doi: 10.6018/analesps.30.3.199361

Medrano, L., & Perez, E. (2019). *Manual de psicometría y evaluación psicológica*. Córdoba: Brujas.

Ministerio de Educación. (2019). *Sistema Especializado en Reporte de Casos de Violencia Escolar*. Perú. Recuperado de http://www.siseve.pe/Web

Nunnally, J. (2013). *Teoría psicométrica.* Ciudad de México: Trillas.

Observatorio Nacional de Política Criminal. (2017). *Adolescentes infractores en el Perú.* Recuperado de https://cdn.www.gob.pe/uploads/document/file/1708343/BOLETIN%2006%20-%202017%20Adolescentes%20Infractores.pdf

Pechorro, P., Ray, J. V., Raine, A., Marocco, J., & Gonçalves, R. A. (2015). The Reactive-Proactive Aggression Questionnaire: Validation among a portuguese sample of incarcerated juvenile delinquents. *Journal of Interpersonal Violence, 32*(13), 1995-2017. doi: 10.1177/0886260515590784

Penado, M., Andreu, J., & Peña, E. (2014). Agresividad reactiva, proactiva y mixta: Análisis de los factores de riesgo individual. *Anuario de Psicología Jurídica, 24*(1), 37-42. doi: 10.1016/j.apj.2014.07.012

Pérez, E., & Medrano, L. (2010). Análisis factorial exploratorio: Bases conceptuales y metodológicas. *Revista Argentina de Ciencias del Comportamiento, 2*(1), 58-66. Recuperado de https://revistas.unc.edu.ar/index.php/racc

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. doi: 10.1016/j.dr.2016.06.004

Raine, A., Dogde, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., … & Liu, J. (2006). The Reactive-Proactive Aggression Questionnaire: Differential correlates of reactive and proactive aggression in adolescent boys. *Aggressive Behavior, 32*(2), 159-171. doi: 10.1002/ab.20115

Ramírez, J. M., & Andreu, J. M. (2006). Aggression, and some related psychological constructs (anger, hostility, and impulsivity). *Neuroscience and Biobehavioral Reviews, 30*(3), 276-291. doi: 10.1016/j.neubiorev.2005.04.015

Rojas-Zegarra, M. E., Arias-Gallegos, W. L., Rivera, R., Geldres-García, J. A., Starke-Moscoso, M. A., & Apaza-Bejarano, E. N. (2020). Propiedades psicométricas de los cuestionarios Reactive/Proactive Questionnaire (RPQ) y How I Think Questionaire

(HIT) en estudiantes peruanos. *Revista de Psicopatología y Psicología Clínica, 25*(1), 59-68. doi: 10.5944/rppc.24426

Romaní, F., Gutiérrez, C., & Lama, M. (2011). Auto-reporte de agresividad escolar y factores asociados en escolares peruanos de educación secundaria. *Revista Peruana de Epidemiología, 15*(2). Recuperado de https://sisbib.unmsm.edu.pe/BVrevistas/epidemiologia/epidemiologia.htm

Roncero, D., Andreu, J. M., & Peña, M. E. (2016). Procesos cognitivos distorsionados en la conducta agresiva y antisocial en adolescentes. *Anuario de Psicología Jurídica, 26*(1), 88-101. doi: 10.1016/j.apj.2016.04.002

Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., … & Du, H. (2021). Lavaan: Latent Variable Analysis (version 0.6-9) [R package]. Recuperado de https://cran.r- project.org/web/packages/lavaan/index.html

RStudio Team. (2015). RStudio: Integrated Development for R (1.4.1717) [Software de cómputo]. Recuperado de http://www.rstudio.com

Ruiz, M. A., Pardo, A., & San Martín, R. (2010). Modelos de ecuaciones estructurales. *Papeles Del Psicólogo, 31*(1), 34-45. Recuperado de http://www.papelesdelpsicologo.es

Torregrosa, M. S., Inglés, C., & García-Fernández, J. M. (2011). El comportamiento agresivo como predictor del autoconcepto: Estudio con una muestra de estudiantes españoles de educación secundaria obligatoria. *Psychosocial Intervention, 20*(2), 201-212. doi: 10.5093/in2011v20n2a8

Torregrosa, M. S., Inglés, C., García-Fernández, J. M., Ruiz-Esteban, C., López-García, K. S., & Zhou, X. (2010). Diferencias en conducta agresiva entre adolescentes españoles, chinos y mexicanos. *European Journal of Education and Psychology, 3*(2), 167-176. doi: 10.30552/ejep.v3i2.41

Torregrosa, M. S., Inglés, C. J., García-Fernández, J. M., Valle, A., & Núñez, J. C. (2012). Relaciones entre conducta agresiva y metas académicas: Estudio con una muestra de estudiantes españoles de educación secundaria obligatoria. *Universitas Psychologica, 11*(4), 1303-1315. doi: 10.11144/Javeriana.upsy11-4.rcam

Vega-Cauich, J. I., & Zumárraga-García, F. M. (2019). Variables asociadas al inicio y consumo actual de sustancias en adolescentes en conflicto con la ley. *Anuario de Psicología Jurídica, 29*(1), 21-29. doi: 10.5093/apj2018a13

Ventura-León, J. (2018a). Intervalos de confianza para el coeficiente omega. Propuesta para el cálculo. *Adicciones, 30*(1), 77-78. Recuperado de https://www.adicciones.es/index.php/adicciones

Ventura-León, J. (2018b). Otras formas de entender la d de Cohen. *Revista Evaluar, 18*(3). doi: 10.35670/1667-4545.v18.n3.22305

Verdam, M. G., Oort, F. J., & Sprangers, M. A. G. (2016). Using structural equation modeling to detect response shifts and true change in discrete variables: An application to the items of the SF-36. *Quality of Life Research, 25*(6), 1361-1383. doi: 10.1007/s11136-015-1195-0

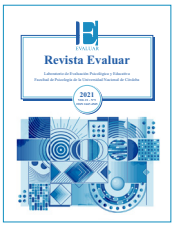
Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3), 1-13. doi: 10.33151/ajp.8.3.93

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014-1045. doi: 10.1007/s11336-016-9506-0

EVALUAR

2021, Vol. 21, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



Propiedades psicométricas del test de temperamento ECBQ-VSF en infantes de diferentes contextos socioeconómicos

Psychometric Properties of the ECBQ-VSF Temperament Test in Infants from Different Socioeconomic Contexts

Lucas G. Gago-Galvagno * [1, 2, 3], Angel M. Elgier [1, 2, 3], Christian Schetsche [1], Carolina De Grandis [1, 2, 3], Florencia Gómez [1], Luis C. Jaume [1, 3], Guadalupe Sosa [1], Susana C. Azzollini [1, 3]

1 - Instituto de Investigaciones, Facultad de Psicología. Universidad de Buenos Aires (UBA). Laboratorio de Cognición y Políticas Públicas. Buenos Aires, Argentina.
2 - Facultad de Psicología y Relaciones Humanas. Universidad Abierta Interamericana. Buenos Aires, Argentina.
3 - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Buenos Aires, Argentina.

Recibido: 25/02/2021 **Revisado:** 03/05/2021 **Aceptado:** 07/05/2021

Resumen

Este trabajo analiza las propiedades psicométricas de la versión en castellano del test de temperamento Early Child Behavior Questionnaire Very Short Form (ECBQ-VSF) en una muestra de 401 infantes (M_{edad} = 27.05 meses DE = 7.08) de Buenos Aires. El análisis factorial exploratorio mostró una estructura de tres factores que explicó el 27% de la varianza, con aceptables pesos factoriales de sus ítems. El análisis factorial confirmatorio indicó que el ajuste del modelo propuesto fue aceptable, siguiendo la estructura factorial original de tres subescalas. El análisis de la confiabilidad señaló una buena consistencia interna para esta muestra. A su vez, esta versión mostró resultados esperables relativos a las correlaciones entre las distintas subdimensiones de la escala, y diferencias según género, nivel socioeconómico y edad. Por último, se encontró invarianza factorial en cuanto al género de los infantes. Se concluye que las diferencias encontradas con respecto a estudios previos podrían deberse a características culturales.

Palabras clave: *temperamento, ECBQ, esfuerzo de control, extraversión, afecto negativo, infantes*

Abstract

This work analyzes the psychometric properties of the Spanish version of the Early Child Behavior Questionnaire Very Short Form temperament test (ECBQ-VSF) in a sample of 401 infants (M_{age} = 27.05 months SD = 7.08) from Buenos Aires. The exploratory factor analysis showed a three-factor structure that explained 27% of the variance, with acceptable factorial weights of its items. The confirmatory factor analysis indicated that the fit of the proposed model was acceptable, following the original factor structure of three subscales. The reliability analysis indicated good internal consistency for this sample. In turn, this version showed expected results related to the correlations between the different sub-dimensions of the scale, and differences according to gender, socioeconomic level and age. Finally, factorial invariance was found regarding the gender of the infants. It is concluded that the differences found with respect to previous studies could be due to cultural characteristics.

Keywords: *temperament, ECBQ, effortful control, extraversion, negative affect, infant*

***Correspondencia a:** Lucas Gago-Galvagno. E-mail: lucas.gagogalvagno@hotmail.com

Cómo citar este artículo: Gago-Galvagno, L. G., Elgier, A. M., Schetsche, C., De Grandis, C., Gómez, F., Jaume, L. C., Sosa, G., & Azzollini, S. C. (2021). Propiedades psicométricas de la escala de temperamento ECBQ-VSF en infantes de diferentes socioeconómicos de Buenos Aires. *Revista Evaluar, 21*(3), 49-62. Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Participaron en la edición de este artículo: Laura Angelelli, Julian Narvaja, Eugenia Maiorana, Eugenia Barrionuevo, Alicia Molinari, Mónica Serppe, Juan Balverdi, Florencia Ruiz, Benjamín Casanova, Ricardo Hernández.

Introducción

Temperamento

El *temperamento* es definido por Rothbart y Bates (2006) como las diferencias individuales en los bebés y niños/as que existen antes de que se desarrollen otros aspectos cognitivos de la personalidad. Este concepto relaciona aspectos genéticos y ambientales que ponen en juego determinadas respuestas afectivas, atencionales y motoras del sujeto según las diferentes demandas del contexto (Rothbart, 2011). También tienen un rol en el despliegue de conductas sociales y en el desempeño social (Calkins, 2005).

Rothbart, Ahadi, Hershey y Fisher (2001) señalan tres factores del temperamento infantil: a) *extraversión*, definida por el nivel de actividad, anticipación positiva, búsqueda de sensaciones, sonrisa-risa, impulsividad y placer intenso. Esta hace referencia al entusiasmo vital, el cual se refleja en el nivel de energía y compromiso con el medio (Squillace-Louhau & Picón-Janeiro, 2017; Yap, Allen, & Sheeber, 2007), b) *esfuerzo de control*, que se define por control inhibitorio, atención focalizada, placer de baja intensidad y sensibilidad perceptiva. El cambio o focalización de la atención permite modular la activación emocional, regulando la exposición a un estímulo y los procesos cognitivos relacionados con esas experiencias emocionales (Rothbart, Ahadi, & Evans, 2000), lo que permite manejar las conductas manifiestas que se asocian con las emociones (Eisenberg, Fabes, Guthrie, & Reiser, 2000) y c) *afectividad negativa*, que comprende malestar, miedo, enojo, tristeza, timidez y dificultad para calmarse. Se refiere a una angustia emocional general y determinada susceptibilidad a emociones negativas (Rothbart et al., 2000), además, se toma como un factor de vulnerabilidad para el desarrollo de trastornos del estado de ánimo y ansiedad (Clark, Watson, & Mineka, 1994).

En investigaciones previas con infantes, se ha encontrado que la extraversión se asocia de forma negativa con el esfuerzo de control y positiva con la afectividad negativa, mientras que el esfuerzo de control se relaciona de forma negativa con la afectividad negativa, aunque los tamaños del efecto fueron bajos ($r < \pm .15$; Putnam, Jacobs, Gartstein, & Rothbart, 2010; Stępień-Nycz, Rostek, Białecka-Pikul, & Białek, 2018). A su vez, debido a las crianzas diferenciales que reciben los infantes según el género, se han encontrado mayores niveles de esfuerzo de control, extraversión y afecto negativo en el género femenino, al recibir mayores niveles de responsividad, interacciones y calidez durante la crianza (Atzaba-Poria & Pike, 2008; Bornstein et al., 2015). También se hallaron diferencias según la edad, con mayores tendencias al esfuerzo de control a medida que los infantes se desarrollan (Hyde, 2014; Reyna & Brussino, 2015), y según el nivel socioeconómico, con menores niveles de esta subescala y mayores de afecto negativo en los infantes pertenecientes a sectores vulnerables (Gago-Galvagno et al., 2019; Segretin, Prats, & Lipina, 2019). Esto último podría deberse a que estos entornos promueven mayores niveles de estrés en cuidadores, estilos parentales autoritarios (Richaud et al., 2013), hacinamiento en la vivienda junto a una mala nutrición (INDEC, 2020; ODSA-UCA, 2020), entre otras situaciones que condicionan las habilidades autorregulatorias infantiles (Gago-Galvagno et al., 2019).

Escala de temperamento ECBQ-VSF

Por otro lado, la escala breve de temperamento Early Child Behavior Questionnaire-Very Short Form (en adelante, ECBQ-VSF) fue creada por Putnam et al. (2010), con el fin de reducir el tiempo de toma que requería la escala original de

201 ítems (una hora aproximadamente; Putnam, Gartstein, & Rothbart, 2006). Se compone de un total de 36 ítems y se representa a cada una de las tres subescalas mencionadas mediante 12 ítems. Los ítems incluidos fueron aquellos que mostraron mayores asociaciones con los factores y menores niveles de correlación entre los otros dos factores (Putnam et al., 2010). Estudios previos han encontrado que el temperamento infantil medido a través del ECBQ-VSF en infantes de 18 a 36 meses se asocia y predice comportamientos de autorregulación y funciones ejecutivas (Epstein et al., 2018; Frick et al., 2018; Gago-Galvagno et al., 2019), habilidades sociales de los infantes (Montenegro & Gago-Galvagno, 2020; Sanson, Hemphill, & Smart, 2004), interacciones tempranas entre cuidador primario e infante (Álvarez, Cristi, Del Real, & Farkas, 2019; Benga, Susa-Erdogan, Friedlmeier, Corapci, & Romonti, 2019), el lenguaje verbal y no verbal (Gago-Galvagno et al., 2019; Salley & Dixon Jr, 2007; Villareal-Garza & Falcón-Albarrán, 2015), entre otras conductas fundamentales para el desarrollo posterior del infante.

En el trabajo original de Putnam et al. (2010) se evaluaron seis muestras estadounidenses de infantes de 18 a 36 meses (n = 488), y se encontraron índices satisfactorios de consistencia interna (.71 en promedio). Además, el instrumento demostró estabilidad longitudinal a los 3 meses (.65 en promedio), aunque con menores niveles a los 12 y 18 meses (.54 y .50 respectivamente); buena validez de criterio (.78), adecuado acuerdo entre los reportes de la madre y el padre (.31), y un buen ajuste en el análisis factorial confirmatorio (AFC; TLI = .977, CFI = .980, RMSEA = .056).

En otra investigación que evaluó las propiedades psicométricas del ECBQ-VSF, utilizando una muestra de infantes estadounidenses de 12 a 36 meses de sectores vulnerables (n = 215), no se encontró un ajuste al modelo mediante AFC (RMSEA = .08, CFI = .46, NFI = .33), las cargas factoriales fueron bajas y complejas y en cuanto a la validez predictiva no se encontraron asociaciones con sobrepeso, aunque se encontraron relaciones negativas con bajo tamaño del efecto (-.17 a .35) entre extraversión y esfuerzo de control con el uso de comida para calmar al infante (James, 2013). Esto podría deberse a que la muestra era de menor edad (desde los 12 meses), por ende, los niveles de autorregulación podrían ser menores. A su vez, el nivel socioeconómico diverso podría jugar en la estabilidad de respuestas intersujeto.

Sin embargo, Potměšilová y Potměšil (2019) sí encontraron la estructura típica en la versión muy breve en una muestra representativa de 709 infantes de 18 a 36 meses de República Checa. Los niveles de alfa fueron de .47 a .77, hubo diferencias por género y edad (aumentando el esfuerzo de control a mayor edad y en el género femenino), aunque fueron descartados cinco ítems por cargar en más de un factor. Los ítems cargaron entre .31 y .60, con un nivel de Kaiser-Meyer-Olkin aceptable (KMO = .76, $p < .01$). A similares resultados se arribó en otro estudio realizado en Polonia con la versión completa del ECBQ (Stępień-Nycz et al., 2018). Este mostró los mismos tres factores que la versión original estadounidense luego de aplicar el análisis factorial exploratorio (AFE), con cargas factoriales entre .35 y .70, una consistencia interna entre .66 y .90 para las 18 subescalas de este instrumento, validez convergente con medidas de laboratorio de temperamento, adecuada estabilidad longitudinal y diferencias por género, evidenciando que los infantes del género femenino puntuaron más alto en las escalas de afecto negativo y esfuerzo de control, tal como se encontró en otros estudios (Putnam et al., 2006). Sin embargo, la varianza explicada fue del 50% (Stępień-Nycz et al., 2018). También Lim, Bae y Lee (2017) trabajaron con

una muestra de infantes de Corea y encontraron los típicos tres factores, con una varianza explicada del 53%, consistencias internas de .63 a .88, validez convergente con la escala EAS, y un correcto ajuste (TLI = .962, CFI = .973, RMSEA = .039 a los 24 meses, y TLI = 1.017, CFI = 1.000, RMSEA = .000 en el grupo de 30 meses).

Por último, en Argentina esta escala fue utilizada en un estudio de Gago-Galvagno et al. (2019) con muestra vulnerable y no vulnerable (n = 60) de infantes argentinos de 18 a 24 meses. La consistencia interna encontrada fue menor a la de estudios previos (.61 a .65), pudiéndose deber esto al rango etario de la muestra evaluada, al tamaño de esta y a los diferentes contextos socioeconómicos evaluados. Este último punto es reforzado en otro estudio (Montenegro & Gago-Galvagno, 2020) que emplea esta escala en Argentina, en una muestra de infantes no vulnerables (n = 102) de 18 a 36 meses, en el que la consistencia interna fue adecuada a alta en comparación con estudios previos (.72 a .81). Sin embargo, en otra investigación realizada con infantes de diferentes contextos socioeconómicos de Chile (n = 90) de 28 a 33 meses, el alfa no se vio modificado con respecto a estudios anteriores (.68 a .76; Álvarez et al., 2019).

Debido a la falta de una adaptación del instrumento ECBQ-VSF en Argentina, a la importancia que posee este instrumento para predecir otros comportamientos en la infancia temprana, a los diferentes índices encontrados en las diversas investigaciones y a los bajos índices de consistencia interna encontrados en muestras de infantes vulnerables de Argentina, el objetivo de la presente investigación consiste en analizar las propiedades psicométricas de la versión en castellano del ECBQ-VSF en una muestra de infantes de Buenos Aires de diferentes contextos socioeconómicos, analizando la estructura factorial, consistencia interna y validez de constructo de esta escala y diferencias por género, edad y nivel socioeconómico. Se espera encontrar, al igual que en la mayoría de las investigaciones, un correcto ajuste al modelo clásico de tres factores (p. ej., extraversión, afecto negativo y esfuerzo de control) y diferencias por variables sociodemográficas en las características temperamentales.

Método

Participantes

Se utilizó un muestreo no probabilístico de tipo intencional y por bola de nieve. Los criterios de selección de la muestra fueron: que los infantes presentaran desarrollo típico, no demostraran antecedentes psiquiátricos o neurológicos y que hubiesen nacido a término. Se evaluaron 401 cuidadores primarios de infantes de 18 a 36 meses (M_{edad} = 27.05 meses, DE = 7.08; femenino = 193), pertenecientes a la Ciudad Autónoma de Buenos Aires y Gran Buenos Aires. De los infantes, 190 (47%) tenían entre 18 y 36 meses, y 211 (53%) tenían entre 27 y 36 meses. A su vez, 245 (61%) asistían a jardines maternales de gestión pública y privada. A su vez, 168 familias presentaron necesidades básicas insatisfechas (NBI; 42%). Las familias eran catalogadas dentro del grupo con NBI si presentaban al menos uno de los siguientes criterios: educación secundaria incompleta, desempleo, vivienda precaria o presencia de hacinamiento.

Instrumentos

Cuestionario sociodemográfico ad-hoc. Se evaluó: a) si los infantes asistían a jardín maternal (*Sí-No*); b) presencia de necesidades básicas insatisfechas (p. ej., educación secundaria incompleta, desempleo, vivienda precaria y presencia

de hacinamiento); *Necesidades básicas satisfechas-Necesidades básicas insatisfechas*; c) género (*masculino-femenino*) y d) edad en meses de los infantes.

Cuestionario de Conducta de Niñez Temprana versión resumida (ECBQ-VSF, versión latinoamericana, Putnam et al., 2010). Evalúa la conducta emocional de los niños desde el punto de vista de los cuidadores. La conducta es clasificada siguiendo una escala Likert de 8 puntos que significa 1) *Nunca*, 2) *Casi nunca*, 3) *Menos de la mitad del tiempo*, 4) *Aproximadamente la mitad del tiempo*, 5) *Más de la mitad del tiempo*, 6) *Casi siempre*, 7) *Siempre* y 8) *No sucedió* (se asignó un 0 en este caso). Esta prueba consta de 36 ítems y cada subescala está formada por 12 ítems. Las subescalas son las de *extraversión*, *afectividad negativa* y *esfuerzo de control*. La escala de esfuerzo de control evalúa la capacidad de inhibir o suprimir las respuestas dominantes. La de extraversión se relaciona con la emoción positiva, el enfoque rápido de recompensas potenciales y un alto nivel de actividad. Finalmente, el afecto negativo incluye predisposición al miedo, ansiedad, tristeza, frustración e incomodidad. El puntaje final de cada una se forma a partir del promedio de los ítems, siendo de 0 a 7 el rango de puntaje.

Procedimiento

Las escalas fueron administradas de forma voluntaria, anónima y confidencial a los cuidadores primarios de los infantes mediante Google Formularios® (n = 215) y de forma presencial (n = 186). Se tomaron mediante estos dos formatos debido a que en sectores vulnerables los investigadores debían estar presentes para asegurarse de que se comprendieran las consignas. Los investigadores proporcionaron mails de contacto en las evaluaciones virtuales, que fueron consultados por los participantes ante cualquier interrogante con respecto a la resolución del cuestionario. Al comienzo de los formularios se explicó a los participantes los objetivos de la investigación y acerca de los ítems del instrumento. Luego debían aceptar su participación y completar los formularios, que fueron aplicados siempre en el mismo orden (sociodemográfico y temperamento) para establecer un control por equiparación y repartir equitativamente el efecto de aprendizaje y fatiga en el total de la muestra. Esta última fue reclutada a través de grupos de Facebook y WhatsApp, contacto con directores/as de jardines maternales, y con figuras de cuidado primario de los infantes, quienes, a su vez, enviaron este formulario a distintos jardines maternales u hogares.

Análisis de datos

Los análisis estadísticos fueron realizados con el software R de Core Team (2020), utilizando un valor de probabilidad de $p < .05$ en todos los casos. Se analizó la normalidad multivariante, mediante el paquete MVN de Korkmaz et al. (2014), con la realización de la prueba de Mardia (1970). Se encontró que la muestra no cumplía con dicho supuesto. Además, mediante el paquete psych de Revelle (2018), se analizaron los estadísticos descriptivos, media, desviación estándar, asimetría y curtosis para todos los ítems de la escala. Se realizó además el análisis factorial exploratorio, el cálculo de las consistencias internas (α y ω), las asociaciones (mediante la prueba parcial Rho de Spearman) y comparaciones de grupo insertando covariables (mediante la prueba MANOVA), y el análisis factorial confirmatorio con el paquete lavaan (Rosseel, 2012).

Tabla 1
Cargas factoriales e índices de complejidad de Hofmann (1978) del EFA EBQ-12.

Variable	Extroversión	Afecto negativo	Esfuerzo de control	ICH
EBQ15	**.589**	-.073	.023	1.034
EBQ8	**.561**	.011	.072	1.034
EBQ13	**.524**	.009	-.056	1.023
EBQ36	**.483**	-.009	-.050	1.022
EBQ17	-.040	**.648**	.054	1.021
EBQ19	-.019	**.631**	-.085	1.038
EBQ16	.042	**.422**	.106	1.147
EBQ23	.004	**.323**	-.063	1.076
EBQ26	.127	.108	**-.651**	1.132
EBQ27	.135	.108	**.648**	1.144
EBQ21	.131	-.024	**.343**	1.295
EBQ14R	.070	-.031	**.247**	1.195

Nota. n = 401; ICH, índice de complejidad de Hofmann (1978).

Resultados

Análisis factorial exploratorio

Como la muestra no cumplía con el supuesto de normalidad multivariante, se realizó un análisis factorial exploratorio siguiendo el método de los ejes principales (Fabrigar et al., 1999), un análisis paralelo y método de rotación oblimin (Costello & Osborne, 2005). La comprobación del *screen plot* justificó la extracción de tres factores, a través de dicha estructura factorial y tras la reducción de ítems, se explicó el 27% de la varianza total del instrumento, lo cual no es un valor aceptable según Merenda (1997). Se dejaron 12 ítems siguiendo el método de observar primero cuál tenía el mayor índice de complejidad de Hofmann (1978), luego eliminando aquel ítem, y por último realizando nuevamente un análisis factorial exploratorio para determinar el siguiente ítem que convendría eliminar. Aunque las cargas factoriales deben ser > .30 según Nunnally (1978), se tomó la decisión de conservar el ítem 14, por los siguientes motivos: a) no mostraba complejidad y por ende el índice de Hofmann (1978) era cercano a 1, b) para no perjudicar las consistencias internas, y c) para que cada dimensión tuviera al menos cuatro ítems.

En la Tabla 1 se observan las cargas factoriales para cada ítem y los índices de complejidad de Hofmann (1978), los cuales son favorables con valores cercanos a 1, ya que esto implica que aportan solo a un factor. Por último, los índices de homogeneidad corregida, los cuales son todos > .15, por lo que son valores adecuados según Hofmann y Hinton (2014).

En la Tabla 2 se observan los ítems que finalmente conforman la escala, y también la estadística descriptiva para cada ítem con la división según las tres dimensiones que representan al temperamento infantil: extraversión, afecto negativo y esfuerzo de control. Además, considerando el índice de límites aceptables de asimetría y curtosis de ± 2 (Hinton, 2014), se puede afirmar que en todos los ítems no hay valores atípicos extremos en la muestra (asimetría máx. = 1.253, curtosis máx. = -1.260).

Tabla 2
Estadística descriptiva de cada ítem de la escala, las frases correspondientes y sus dimensiones.

Dimensión	Ítem	Frase	M(DE)	Asimetría	Curtosis
E	15	Durante las actividades diarias, ¿con qué frecuencia prestó atención de forma correcta cuando usted le llamó?	5.32 (1.68)	-0.827	-0.366
	8	Cuando se concentró en su juguete o juego favorito, al mismo tiempo que jugaba, ¿atendió a lo que se le preguntó?	4.53 (2.06)	-0.359	-1.260
	13	Cuando descubrió una nueva actividad, ¿con qué frecuencia se integró al juego rápidamente?	5.64 (1.44)	-1.224	1.031
	36	Al estar en reuniones familiares con adultos o niños, ¿con qué frecuencia disfrutó jugando con personas distintas?	5.08 (1.93)	-0.690	-0.820
AN	17	Durante las actividades diarias, ¿le molestaron los sonidos de lugares ruidosos?	2.60 (2.07)	1.066	-0.324
	19	En un lugar público, ¿con qué frecuencia pareció asustado por los vehículos grandes o ruidosos?	2.43 (1.97)	1.253	0.148
	16	Durante las actividades diarias, ¿pareció estar molesto por las etiquetas en su ropa?	2.59 (2.30)	1.095	-0.520
	23	Después de una actividad o acontecimiento emocionante, ¿con qué frecuencia pareció desanimado o melancólico?	2.56 (1.87)	1.134	0.093
EC	26	Cuando pidió algo y usted le dijo "no", ¿con qué frecuencia hizo un berrinche?	5.11 (1.84)	-0.733	-0.632
	27	Cuando le pidió esperar para algo que deseaba (como una golosina), ¿con qué frecuencia esperó pacientemente?	3.35 (1.99)	0.442	-1.077
	21	Cuando le pidió que "no" realizara alguna actividad, ¿con qué frecuencia interrumpió una actividad rápidamente?	3.53 (1.85)	0.449	-0.905
	14R	Cuando se involucró en una actividad que requería poner atención (como construir con bloques, armar un rompecabezas o vestir a una muñeca) ¿con qué frecuencia se cansó de la actividad relativamente rápido?	4.13 (1.90)	-0.167	-1.173

Nota. n = 401; E = Extraversión; AN = Afectividad negativa; EC = Esfuerzo de control. Todas las frases son referidas al infante.

Tabla 3
Consistencias internas de la escala.

	Extraversión	Afecto negativo	Esfuerzo de control	Media de α y ω
Alfa de Cronbach (α)	.620	.533	.566	.573
Coeficiente omega (ω)	.676	.590	.660	.642

A su vez, se calcularon las consistencias internas (Tabla 3). Tanto el coeficiente alfa de Cronbach (α), como el coeficiente omega (ω) se observan dentro de los rangos adecuados (Hinton, 2014; Katz, 2006).

Asociaciones entre variables

Por otro lado, utilizando la prueba Rho de Spearman parcial, controlando por edad, género y nivel socioeconómico de los infantes, se encontraron asociaciones entre los tres subdimensiones. El *esfuerzo de control* se asoció de forma negativa con la *afectividad negativa* (Rho = -.183; p = .001) y positiva con la *extraversión* (Rho = .236, p = .001). No se encontraron asociaciones entre la *extraversión* y la *afectividad negativa* (p > .05).

También se encontraron diferencias entre los grupos por género, insertando la variable edad en meses y nivel socioeconómico de los infantes como variables de control. El género femenino mostró mayores niveles de *extraversión* (F = 7.42, p = .007, eta^2 = .019) y menores de *afecto negativo* (F = 5.76, p = .017, eta^2 = .014) en comparación con los del género masculino.

En cuanto a los infantes de sectores vulnerables, controlando por género y edad de los infantes en meses, se encontraron mayores niveles de *extraversión* (F = 4.56, p = .034, eta^2 = .009) y *afecto negativo* (F = 8.46, p = .004, eta^2 = .021) en comparación con los infantes de sectores no vulnerables.

Sobre las respuestas en formato virtual y presencial, solo se encontraron mayores niveles de *extraversión* en el grupo que respondió de forma presencial (F = 20.51, p = .001, eta^2 = .049) luego de controlar edad y género.

Por último, en cuanto a la edad y controlando por género y nivel socioeconómico, se encontraron mayores niveles de *esfuerzo de control* (F = 9.93, p = .002, eta^2 = .022) y afecto negativo (F = 6.73, p = .009, eta^2 = .017) en el grupo de 27 a 36 con respecto al de 18 a 26 meses.

Análisis factorial confirmatorio

Por último, según las indicaciones de Hu y Bentler (1999), podemos considerar un modelo como adecuado cuando su ajuste toma los siguientes valores: $\chi^2/gl \leq 3$, SRMR ≤ .08, RMSEA ≤ .06, CFI ≥ .95, TLI ≥ .95. Tras la reducción de ítems derivados del AFE, se obtuvieron los siguientes valores de ajuste: χ^2 DWLS (Mîndrilă, 2010, Chi-cuadrado utilizando los mínimos cuadrados ponderados en diagonal) = 62.125; gl = 51; Scaling (Factor de corrección para el Chi-cuadrado utilizando los mínimos cuadrados ponderados en diagonal) = 1.035; p = .137; RMSEA = .023, 90% CI [.000, .042]; SRMR = .048; CFI = .974 y TLI = .966. Dados estos resultados, el ajuste al modelo puede considerarse adecuado.

Invarianza factorial

Con el fin de analizar si el modelo presenta invarianza factorial, se realizó un análisis de múltiples grupos con relación a los géneros femenino (n = 194) y masculino (n = 207). Debido a la sensibilidad de la prueba de diferencia de Chi-cuadrado en muestras de gran tamaño (Meade et al., 2008), se utilizó el cambio del CFI para poder evaluar la invarianza factorial. Este se encontró dentro de unos rangos adecuados con ΔCFI ≥ -.01 según Cheung y Rensvold (2002). No existió diferencia entre los modelos configural (M1), débil (M2), fuerte (M3) y estricto (M4).

Conclusiones

El objetivo de la presente investigación fue evaluar las propiedades psicométricas de la versión en castellano del ECBQ-VSF en una muestra de infantes de Buenos Aires, analizando la estructura factorial, consistencia interna, validez de constructo de esta escala y diferencias por variables sociodemográficas. Se encontró un correcto ajuste al modelo clásico de tres factores (p. ej., extraversión, afecto negativo y esfuerzo de control), una buena consistencia interna, un bajo nivel de varianza explicada, y asociaciones esperables en cuanto a las subdimensiones del temperamento, género, edad y nivel socioeconómico. Por último, el instrumento demostró invarianza factorial a partir del género.

En esta muestra, se encontró un correcto ajuste con el modelo clásico de tres factores propuesto por los autores originales (Putnam et al., 2010). Esto replica resultados anteriores, pero ahora en una muestra latinoamericana de diversos sectores socioeconómicos, lo cual demuestra la consistencia de la teoría clásica de tres factores del temperamento.

Por otro lado, si bien la consistencia interna fue adecuada, fue algo menor a la obtenida en estudios previos, al igual que la varianza explicada (Lim et al., 2017; Putnam et al., 2010; Stępień-Nycz et al., 2018), aunque similar a la obtenida por James (2013). Si bien en estos estudios ambos índices fueron medios a bajos, los niveles aún menores encontrados en el presente estudio podrían deberse a la muestra de personas de sectores vulnerables y a la toma presencial, lo que podría sesgar los resultados y producir respuestas por aquiescencia o deseabilidad social. Esto se observa en los mayores niveles de extroversión encontrados en los sectores vulnerables y en formato de toma presencial, ya que en presencia del evaluador el sujeto podría sobrevalorar características de los infantes que son funcionales a la cultura en donde vive (Putnam et al., 2006; Rothbart & Bates, 2006). A su vez, si bien las personas de sectores vulnerables contaban con la presencia de los investigadores, podrían no haber comprendido las consignas debido a la falta de habituación a responder este tipo de cuestionarios. Esta muestra, a su vez, es significativamente disímil a la utilizada para desarrollar el ECBQ-VSF, no solo por cuestiones socioeconómicas, si no también culturales (ver Krassner et al., 2017), lo que podría contribuir a estas diferencias halladas en los índices.

En cuanto a las asociaciones entre las subdimensiones, replican resultados anteriores (Putnam et al., 2010; Stępień-Nycz et al., 2018), y evidencian que el *esfuerzo de control* se asoció de forma negativa con la *afectividad negativa* y positiva con la *extraversión*. El *esfuerzo de control* refiere a la capacidad de inhibir respuestas preponderantes, lo cual disminuye los niveles de *arousal* negativo y promueve en ocasiones la socialización (Rothbart & Bates, 2006; Rothbart, 2011). Sin embargo, es necesario resaltar que no se encontraron asociaciones entre la *extraversión* y la *afectividad negativa*, que suelen relacionarse de forma positiva (Calkins, 2005; Rothbart et al., 2001). Esto podría deberse a que en Argentina y algunos países de América los niveles generales de socialización, *afecto negativo* y *extraversión* en infantes suelen ser en general más altos que en Europa y Asia (Gago-Galvagno et al., 2019; Krassner et al., 2017), lo cual podría disminuir el grado de asociación entre estas variables.

A su vez, se encontraron los resultados y tamaños del efecto esperables en cuanto al género, los cuales informaron que los infantes del género femenino mostraban mayores niveles de *extraversión* y *afecto negativo* (Atzaba-Poria & Pike, 2008; Bornstein et al., 2008). Esto suele interpretarse como un resultado del trato diferenciado

que demuestran los cuidadores con los infantes del género femenino, en donde en general hay un aspecto más emocional, cálido e interactivo ligado a la comunicación temprana con respecto a los infantes masculinos (Bornstein et al., 2015). Sin embargo, no se encontraron los resultados típicos en cuanto a mayores niveles de *esfuerzo de control* por parte de este género (Else-Quest, Hyde, Goldsmith, & Van Hulle, 2006; Potměšilová & Potměšil, 2019; Reyna & Brussino, 2015). Esto puede deberse a las diferencias susodichas con respecto al tipo de muestra evaluada, aunque otros estudios también encontraron estabilidad en ambos géneros en cuanto a sus características temperamentales (Bornstein et al., 2015; Hyde, 2014). Más investigaciones deberían llevarse a cabo en diferentes culturas para subsanar estas contradicciones.

En cuanto a la edad, los resultados y tamaños del efecto son consistentes en lo relativo a mayores niveles de esfuerzo de control en infantes de mayor edad (Hyde, 2014; Reyna & Brussino, 2015), y muestran que a medida que el infante se desarrolla, posee más habilidades de autocontrol. Sin embargo, la *afectividad negativa* también fue mayor en el grupo de mayor rango etario, al contrario de otras investigaciones (Bornstein et al., 2015; Reyna & Brussino, 2015) pero de modo similar a los resultados de otros estudios (Gago-Galvagno et al., 2019; Stępień-Nycz et al., 2018). Esta inconsistencia en los resultados podría deberse a que en este período aún se encuentran en desarrollo las habilidades de control de impulsos emocionales (Rothbart & Bates, 2006; Rothbart, 2011), y son relativamente dependientes de los niveles de autorregulación de los cuidadores primarios (Calkins, 2005; Rothbart & Bates, 2006). Este punto también podría demostrarse en los menores índices de consistencia interna encontrados en estudios que utilizaron muestras de infantes mayores a 18 meses (Álvarez et al., 2019; James, 2013).

Por último, los resultados en cuanto al nivel socioeconómico que presentan mayores niveles de *extraversión* y *afecto negativo* en infantes de sectores vulnerables son congruentes con otras investigaciones en temperamento (Gago-Galvagno et al., 2019; Segretin et al., 2019) y en regulación emocional y cognitiva (Brandes-Aitken, Braren, Swingler, Voegtline, & Blair, 2019; Gago Galvagno et al., 2019), en donde se encontraron menores niveles de regulación en infantes provenientes de sectores vulnerables. Esto deviene de un contexto que genera dificultades en las interacciones cuidador-infante y por ende en las habilidades regulatorias, ya que aumentan los niveles de estrés, violencia y hacinamiento, entre muchos otros (INDEC, 2020; ODSA-UCA, 2020).

Si bien la siguiente investigación replica resultados previos y demuestra un correcto ajuste al modelo para una muestra de infantes argentinos de diferentes niveles socioeconómicos, es necesario resaltar una serie de limitaciones. Por un lado, el diseño del estudio es de corte transversal, lo que no permite clarificar el curso temporal de las asociaciones halladas y limita el análisis de las diferencias individuales tan importantes en el estudio del temperamento. Por ende, los puntajes en ciertos factores podrían ser un reflejo de la transición en el desarrollo temperamental, lo cual podría variar si la toma se ejecutara en edades anteriores o posteriores. A su vez, la muestra fue obtenida por técnicas de muestreo no probabilísticas, lo que limita la generalización de los resultados. Por último, el método de recolección virtual podría aumentar la varianza secundaria y por ende de error en la medición, ya que el investigador no está presente durante la toma.

Para futuras investigaciones se espera poder realizar estudios longitudinales para evaluar trayectorias del desarrollo, trabajar con muestreos por cuotas o probabilísticos limitados a una po-

blación objetivo accesible, y realizar la toma en formato presencial en su totalidad, siendo que los reportes parentales, al ser observaciones comportamentales de terceros, pueden traer más confusiones a la hora de ser completados (Carranza-Carnicero, Pérez-López, Gonzáles-Salinas, & Martínez-Fuentes, 2000; Rothbart & Hwang, 2002). De este modo se espera establecer un análisis más acabado de esta variable tan importante en el desarrollo de los infantes y las interacciones tempranas.

Referencias

Álvarez, C., Cristi, P., del Real, M. T., & Farkas, C. (2019). Mentalization in Chilean mothers with children aged 12 and 30 months: Relation to child sex and temperament and family socioeconomic status. *Journal of Child and Family Studies, 28*(4), 959-970. doi: 10.1007/s10826-019-01348-1

Atzaba-Poria, N., & Pike, A. (2008). Correlates of parenting for mothers and fathers from English and Indian backgrounds. *Parenting: Science and Practice, 8*(1), 17-40. doi: 10.1080/15295190701665698

Benga, O., Susa-Erdogan, G., Friedlmeier, W., Corapci, F., & Romonti, M. (2019). Maternal self-construal, maternal socialization of emotions and child emotion regulation in a sample of Romanian mother-toddler dyads. *Frontiers in Psychology, 9*, Article 2680. doi: 10.3389/fpsyg.2018.02680

Bornstein, M. H., Putnick, D. L., Gartstein, M. A., Hahn, C. S., Auestad, N., & O'Connor, D. L. (2015). Infant temperament: Stability by age, gender, birth order, term status, and socioeconomic status. *Child Development, 86*(3), 844-863. doi: 10.1111/cdev.12367

Bornstein, M. H., Putnick, D. L., Heslington, M., Gini, M., Suwalsky, J. T. D., Venuti, P., ... & Zingman de Galperín, C. (2008). Mother-child emotional availability in ecological perspective: Three countries, two regions, two genders. *Developmental Psychology, 44*(3), 666-680. doi: 10.1037/0012-1649.44.3.666

Brandes-Aitken, A., Braren, S., Swingler, M., Voegtline, K., & Blair, C. (2019). Sustained attention in infancy: A foundation for the development of multiple aspects of self-regulation for children in poverty. *Journal of Experimental Child Psychology, 184*, 192-209. doi: 10.1016/j.jecp.2019.04.006

Calkins, S. D. (2005). El temperamento y su impacto en el desarrollo infantil: Comentarios sobre Rothbart, Kagan y Eisenberg. En R. E. Tremblay, M. Boivin & R. Peters (Eds.), *Enciclopedia sobre el desarrollo de la primera infancia*. Nueva York, NY: Wiley. Recuperado de https://www.enciclopedia-infantes.com/temperamento/segun-los-expertos/el-temperamento-y-su-impacto-en-el-desarrollo-infantil-comentarios

Carranza-Carnicero, J. A., Pérez-López, J., González-Salinas, M. del C., & Martínez-Fuentes, M. T. (2000). A longitudinal study of temperament in infancy: Stability and convergence of measures. *European Journal of Personality, 14*(1), 21-37. doi: 10.1002/(sici)1099-0984(200001/02)14:1%3C21::aid-per367%3E3.3.co;2-1

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Clark, L. A., Watson, D., & Mineka, S. (1994). Temperament, personality, and the mood and anxiety disorders. *Journal of Abnormal Psychology, 103*(1), 103-116. doi: 10.1037/0021-843x.103.1.103

Core Team (2020). *R: A language and environment for statistical computing*. https://www.r-project.org

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), 7-17. doi: 10.7275/jyj1-4868

Eisenberg, N., Fabes, R. A., Guthrie, I. K., & Reiser, M. (2000). Dispositional emotionality and regulation: Their role in predicting quality of social functioning.

Journal of Personality and Social Psychology, 78(1), 136-157. doi: 10.1037/0022-3514.78.1.136

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin, 132*(1), 33-72. doi: 10.1037/0033-2909.132.1.33

Epstein, A., Pesce, C., Errázuriz, C., Gómez-Barris, I., Izquierdo, V., & Farkas, C. (2018). Relación de la autorregulación infantil con sensibilidad materna y contexto familiar a los 12 y 30 meses de edad. *Summa Psicológica, 15*(1), 25-34. doi: 10.18774/summa-vol15.num1-360

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272. doi: 10.1037/1082-989X.4.3.272

Frick, M. A., Forslund, T., Fransson, M., Johansson, M., Bohlin, G., & Brocki, K. C. (2018). The role of sustained attention, maternal sensitivity, and infant temperament in the development of early self-regulation. *British Journal of Psychology, 109*(2), 277-298. doi: 10.1111/bjop.12266

Gago-Galvagno, L. G., De Grandis, M. C., Clerici, G. D., Mustaca, A. E., Miller, S. E., & Elgier, A. M. (2019). Regulation during the second year: Executive function and emotion regulation links to joint attention, temperament and social vulnerability in a Latin American sample. *Frontiers in Psychology, 10,* Article 1473. doi: 10.3389/fpsyg.2019.01473

Hinton, P. R. (2014). *Statistics explained*. Routledge.

Hofmann, R. J. (1977). Indices descriptive of factor complexity. *The Journal of General Psychology, 96*(1), 103-110. doi: 10.1080/00221309.1977.9920803

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review Psychology, 65*(1), 373-398. doi: 10.1146/annurev-psych-010213-115057

INDEC. (2020). *Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos*. *Informe Técnico 181.* Buenos Aires. Recuperado de https://www.indec.gob.ar

James, B. L. (2013). *Psychometric evaluation of the early childhood behavior questionnaire very short form in low-income WIC mothers and toddlers* (Master thesis). Pennsylvania State University, USA.

Katz, M. H. (2006). *Multivariable analysis* (2ª ed.). New York, NY: Cambridge University Press.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal, 6*(2), 151-162. doi: 10.32614/RJ-2014-031.

Krassner, A. M., Gartstein, M. A., Park, C., Dragan, W. Ł., Lecannelier, F., & Putnam, S. P. (2017). East-west, collectivist-individualist: A cross-cultural examination of temperament in toddlers from Chile, Poland, South Korea, and the US. *European Journal of Developmental Psychology, 14*(4), 449-464. doi: 10.1080/17405629.2016.1236722

Lim, J. Y., Bae, Y. J., & Lee, Y. J. (2017). Validation study of the Korean version of Rothbart's Early Childhood Behavior Questionnaire. *Korean Journal of Child Studies, 38*(4), 33-47. doi: 10.5723/kjcs.2017.38.4.33

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530. doi: 10.1093/biomet/57.3.519

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592. doi: 10.1037/0021-9010.93.3.568

Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development, 30*(3), 156-164. doi: 10.1080/07481756.1997.12068936

Mîndrilă, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias

with ordinal and multivariate non-normal data. *International Journal for Digital Society, 1*(1), 60-66. doi: 10.20533/ijds.2040.2570.2010.0010

Montenegro, F., & Gago-Galvagno, L. G. (2020). ¿Se relacionan el temperamento, la asistencia a los jardines maternales y el género con las habilidades sociales durante los primeros años de vida? *Revista de Psicología, 19*(2), 107-121. doi: 10.24215/2422572xe066

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

ODSA-UCA. (2020). *Balance general: Deterioros históricos y desigualdades estructurales en el contexto COVID-19*. Informe de avance. Buenos Aires. Recuperado de http://uca.edu.ar/es/observatorio-de-la-deuda-social-argentina

Potměšilová, P., & Potměšil, M. (2019). The Early Childhood Behavior Questionnaire Very Short Form (ECBQ VSF) and its adaptation to the population of the Czech Republic. *Psychiatria i Psychologia Kliniczna, 19*(3), 281-287. doi: 10.15557/pipk.2019.0029

Putnam, S. P., Gartstein, M. A, & Rothbart, M. K. (2006). Measurement of fine-grained aspects of toddler temperament: The Early Childhood Behavior Questionnaire. *Infant Behavior & Development, 29*(3), 386-401. doi: 10.1016/j.infbeh.2006.01.004

Putnam, S. P., Jacobs, J., Gartstein, M. A., & Rothbart, M. K. (2010). *Development and assessment of short and very short forms of the Early Childhood Behavior Questionnaire*. Poster presentado en la International Conference on Infant Studies, Baltimore, MD. Recuperado de https://www.research.bowdoin.edu

Revelle, W. (2018). *Psych: Procedures for personality and psychological research*. Northwestern University.

Reyna, C., & Brussino, S. (2015). Diferencias de edad y género en comportamiento social, temperamento y regulación emocional en niños argentinos. *Acta Colombiana de Psicología, 18*(2), 51-64. doi: 10.14718/acp.2015.18.2.5

Richaud, M. C., Mestre, M. V., Lemos, V. N., Tur, A., Ghiglione, M. E., & Samper, P. (2013). La influencia de la cultura en los estilos parentales en contextos de vulnerabilidad social. *Avances en Psicología Latinoamericana, 31*(2), 419-431. Recuperado de https://revistas.urosario.edu.co/index.php/apl

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. doi: 10.18637/jss.v048.i02

Rothbart, M. (2011). *Becoming who we are. Temperament and personality in development*. New York, NY: The Guilford Press.

Rothbart, M. K. & Bates, J. E. (2006). Temperament. En W. Damon & R. Lerner (Eds.), *Handbook of Child Psychology. Vol 3: Social, emotional, and personality development* (pp. 99-166). New York, NY: Wiley.

Rothbart, M. K., & Hwang, J. (2002). Measuring infant temperament. *Infant Behavior and Development, 25*(1), 113-116. doi: 10.1016/s0163-6383(02)00109-1

Rothbart, M. K., Ahadi, S. A., & Evans, D. E. (2000). Temperament and personality: Origins and outcomes. *Journal of Personality and Social Psychology, 78*(1), 122-135. doi: 10.1037/0022-3514.78.1.122

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72(*5), 1394-1408. doi: 10.1111/1467-8624.00355

Salley, B. J., & Dixon Jr, W. E. (2007). Temperamental and joint attentional predictors of language development. *Merrill-Palmer Quarterly, 53*(1), 131-154. doi: 10.1353/mpq.2007.0004

Sanson, A., Hemphill, S. A., & Smart, D. (2004). Connections between temperament and social development: A review. *Social Development, 13*(1), 142-170. doi: 10.1046/j.1467-9507.2004.00261.x

Segretin, M. S., Prats, L. M., & Lipina, S. J. (2019). Asociaciones entre el temperamento y la regulación del cortisol en preescolares de hogares pobres. *Cuadernos de Neuropsicología/Panamerican Journal of Neuropsychology, 13*(2), 73-91. Recuperado de https://www.cnps.cl/index.php/cnps

Squillace, M., & Picón-Janeiro, J. (2017). Impulsividad,

un constructo multifacético: Validación del CUBI. *Revista Evaluar, 17*(1), 1-17. doi: 10.35670/1667-4545.v17.n1.17070

Stępień-Nycz, M., Rostek, I., Białecka-Pikul, M., & Białek, A. (2018). The Polish adaptation of the Early Childhood Behavior Questionnaire (ECBQ): Psychometric properties, age and gender differences and convergence between the questionnaire and the observational data. *European Journal of Developmental Psychology, 15*(2), 192-213. doi: 10.1080/17405629.2017.1292906

Villareal-Garza, M. A., & Falcón-Albarrán, A. J. (2015). Análisis de la relación entre el temperamento y el aprendizaje léxico. *Investigación y Práctica en Psicología del Desarrollo, 1*, 23-30. doi: 10.33064/ippd1672

Yap, M. B., Allen, N. B., & Sheeber, L. (2007). Using an emotion regulation framework to understand the role of temperament and family processes in risk for adolescent depressive disorders. *Clinical Child and Family Psychology Review, 10*(2), 180-196. doi: 10.1007/s10567-006-0014-0

2021, Vol. 21, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

Propiedades psicométricas de la Escala de Flow Disposicional-2 en videojuegos

The Psychometric Properties of Dispositional Flow Scale-2 in Video Games

Raúl Rodríguez-Antonio * [1] , Jair Arody del Valle-López [1]

1 - Universidad de Montemorelos. Montemorelos, Nuevo León, México.

Recibido: 01/09/2021 **Revisado:** 10/10/2021 **Aceptado:** 25/10/2021

Resumen

El estado de flow es una característica psicológica importante en el contexto del diseño y evaluación de videojuegos educativos. En este estudio se analizaron las propiedades psicométricas de una adaptación mexicana de la Escala de Flow Disposicional-2 en el contexto de los videojuegos. Con base en la información suministrada por una muestra de 312 estudiantes de una universidad del noreste de México, con edades de 16 a 34 años (M = 19.90, DE = 2.73), se realizó un análisis factorial confirmatorio que sugirió un ajuste aceptable de la estructura factorial, con adecuada validez convergente pero deficiente validez discriminante. Adicionalmente, con base en un análisis factorial exploratorio, se identificó un modelo reespecificado que agrupó 33 de los 36 ítems de la escala. Esta estructura factorial, que mostró un ajuste aceptable, y adecuada validez convergente y discriminante, sugiere que las dimensiones de la escala pueden agruparse en antecedentes y consecuencias del flow.

Palabras clave: *análisis factorial exploratorio, análisis factorial confirmatorio, validez convergente, validez discriminante, estudiantes mexicanos, estado de flow, videojuegos*

Abstract

The state of flow is an important psychological characteristic of educational video games design and evaluation. This study analyzed a Mexican adaptation of the Dispositional Flow Scale-2 psychometric properties in the use of video games. Based on the information provided by a sample of 312 students, aged 16 to 34 years (M = 19.90, SD = 2.73), from a university in northeastern Mexico a confirmatory factor analysis that suggested an acceptable fit of the factorial structure, adequate convergent validity but poor discriminant validity was performed. Based on an exploratory factor analysis a re-specified model was identified, grouping 33 of the 36 items of the scale. This factorial structure, which showed an acceptable fit, adequate convergent validity and discriminant validity, suggests that scale dimensions can be grouped into antecedents and consequences of flow.

Keywords: *exploratory factor analysis, confirmatory factor analysis, convergent validity, discriminant validity, Mexican students, flow state, video games*

***Correspondencia a:** Raúl Rodríguez Antonio. Facultad de Educación. Universidad de Montemorelos. Libertad 1300, Pte. Montemorelos, Nuevo León, México. C.P. 67530. Teléfono: +52-8261089023. ORCID: 0000-0001-6766-4133. E-mail: rrodriguez@um.edu.mx

Cómo citar este artículo: Rodríguez-Antonio, R., & Del Valle-López, J. A. (2021). Propiedades psicométricas de la Escala de Flow Disposicional-2 en videojuegos. *Revista Evaluar, 21*(3), 63-80. Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Nota de autor: Los autores declaran no tener ningún conflicto de intereses.

Participaron en la edición de este artículo: Julian Narvaja, Stefano Macri, Eugenia Barrionuevo, Alicia Molinari, Mónica Serppe, Ricardo Hernández.

Introducción

En la actualidad, jugar videojuegos es una de las actividades más atractivas para adolescentes y adultos jóvenes debido a la combinación de diversos elementos tales como desafíos, metas, reglas, resolución de problemas, retroalimentación, emoción y diversión, entre otros, los cuales requieren que el jugador se involucre en un proceso de aprendizaje continuo para lograr el dominio del juego (Domínguez et al., 2013; Prensky, 2001). Varios de los elementos mencionados están presentes en la teoría del flow, la cual ha llegado a ser un elemento fundamental en el diseño de videojuegos educativos (Khoshnoud, Alvarez-Igarzábal, & Wittmann, 2020; Montes-González, Ochoa-Angrino, Baldeón-Padilla, & Bonilla-Sáenz, 2018). Además, estos elementos facilitan el involucramiento activo del estudiante en el aprendizaje y el logro de objetivos educacionales de forma efectiva (Kiili, de Freitas, Arnab, & Lainema, 2012).

La teoría del flow fue desarrollada por Csikszentmihalyi (1975) al estudiar grupos de individuos que desarrollaban actividades autotélicas diversas en un contexto deportivo. Según el autor, una actividad autotélica es aquella que requiere un gasto formal y extensivo de energía por parte del sujeto, aunque le produzca poca o nula recompensa convencional, de modo que el solo acto de llevar a cabo la actividad es la recompensa. En este contexto, se entiende por recompensa convencional a las recompensas extrínsecas, tales como poder, prestigio y reconocimientos, entre otras.

El estado de flow, también llamado experiencia óptima, es "el estado dinámico de una persona que le genera una sensación holística al actuar con un involucramiento total" (Csikszentmihalyi, 1975, p. 36), y tiende a ocurrir cuando una persona enfrenta un conjunto definido de metas que requieren respuestas apropiadas, o cuando una persona emplea intensamente sus habilidades para vencer un desafío muy atractivo que está al alcance de sus posibilidades. El flow se ha conceptualizado con base en nueve dimensiones que a continuación se describen.

La primera dimensión del flow corresponde al *balance habilidad-desafío* (HD). Para que el estado de flow suceda es necesario que la actividad que se realiza presente un desafío que pueda ser cumplido por el individuo con las habilidades que posee. El desafío se refiere a la experiencia subjetiva del individuo derivada de la congruencia entre sus propias habilidades y las demandas de la actividad (Rodríguez-Ardura & Meseguer-Artola, 2017). Un desafío alcanzable fomenta la motivación de la persona, en tanto que uno demasiado difícil o demasiado fácil causa el efecto contrario. En relación a esto último, si las habilidades del sujeto sobrepasan al desafío, el resultado es aburrimiento. Si por el contrario el desafío sobrepasa las habilidades, el sujeto experimenta ansiedad. Si ambos, desafío y habilidades, están debajo del promedio, se presenta apatía (Stavrou & Zervas, 2004).

La dimensión del flow denominada *mezcla de conciencia-acción* (CA) es probablemente uno de los más claros signos de que una persona está experimentando el flow (Csikszentmihalyi, 2014). Cuando una persona está en estado de flow está consciente de sus acciones, mas no de sí misma. Es decir, para que el estado de flow se mantenga, el individuo no puede reflexionar sobre sí mismo. Por ejemplo, al preguntarse: *¿Qué estoy haciendo aquí? ¿Lo estaré haciendo bien?* Si estas reflexiones tienen lugar, el estado de flow se interrumpe (Csikszentmihalyi, 1975). Así, la mezcla de conciencia-acción suele ocurrir en episodios breves que son interrumpidos por el propio actor cuando adopta una perspectiva externa acerca de la acción que está desarrollando.

Otra dimensión del flow corresponde al *establecimiento de metas claras* (MC). Este aspecto se refiere a la formulación de objetivos claros y precisos, que es fundamental para que suceda el estado de flow. Es decir, si se conoce claramente lo que se debe lograr, el individuo se prepara psicológicamente para ejecutar las acciones necesarias con el propósito de alcanzar la meta (Calero & Injoque-Ricle, 2013; Csikszentmihalyi, 2014).

En el estado de flow se presentan demandas de acciones coherentes y no contradictorias que proporcionan al individuo una retroalimentación clara y sin ambigüedades de forma automática en un esquema de acción y reacción, dado que el individuo está tan involucrado en la actividad que no puede reflexionar sobre ella (Csikszentmihalyi, 1975). A esta dimensión se le denomina *retroalimentación no ambigua* (RNA). Así, el flow se experimenta con más frecuencia en actividades que implican reglas de acción definidas, por ejemplo, juegos, procedimientos, rituales o arte.

Cuando un individuo experimenta el estado de flow, su atención responde solo a una limitada cantidad de estímulos. Los elementos motivacionales y las reglas que se asocian con las tareas en el estado de flow parecen definir qué estímulos son relevantes para el individuo y cuáles no lo son (Csikszentmihalyi, 1975). A esta dimensión del flow se la conoce como *concentración en la tarea* (CT), y se evidencia cuando la persona está completamente enfocada en la tarea que está realizando (Giasiranis & Sofos, 2017).

Cuando una persona experimenta el estado de flow está en control de sus habilidades y acciones, lo que le permite cumplir con las demandas del ambiente aún en situaciones que implican cierto riesgo o peligro. A esta dimensión del flow se la conoce como *sentido de control* (SC). Este sentido no necesariamente se presenta de manera consciente u objetiva; más bien el individuo no suele preocuparse por la posible pérdida de control de la situación. El sentido de control, aunque en ciertos escenarios de juego proviene de vencer a un contrincante, se entiende generalmente como un estado de victoria contra las limitaciones del propio individuo (Csikszentmihalyi, 1975).

Otra dimensión del flow es la *pérdida de conciencia* (PC), la cual no significa que el individuo no esté consciente de su cuerpo o sus funciones, o que pierda el contacto con su entorno físico; más bien implica que su propio ego llega a ser irrelevante (Csikszentmihalyi, 1975). El ego es un elemento básico para la vida social, ya que la consciencia del yo permite al individuo realizar las interacciones y negociaciones con otras personas. Dado que las tareas que propician el estado de flow generalmente implican reglas que son aceptadas libremente por el sujeto, éste no se preocupa por decidir qué hacer o qué no hacer, así que no necesita utilizar su yo para efectuar negociaciones sociales (Csikszentmihalyi, 2014).

En el estado de flow, debido a la intensidad de la experiencia, el individuo puede percibir que el espacio temporal se distorsiona (Calero & Injoque-Ricle, 2013; Csikszentmihalyi, 2014). En algunos casos pareciera que los intervalos de tiempo se extienden o se reducen, e inclusive pudiera parecer que el tiempo se detiene. Esta dimensión, conocida como *transformación del tiempo* (TT), se asocia con la liberación de la restricción temporal para que el sujeto cumpla con el desafío (Csikszentmihalyi, 1975).

Por último, la dimensión del flow denominada *experiencia autotélica* (EA) se refiere a realizar una actividad simplemente porque se desea llevarla a cabo sin necesidad de recompensas u objetivos externos. En el estado de flow, el solo acto de realizar una actividad justifica la actividad, ya que "el propósito del flow es mantenerse en el estado de flow" (Csikszentmihalyi, 1975, p. 47).

En el contexto del proceso de enseñanza

aprendizaje el flow ha sido estudiado principalmente en actividades vinculadas con videojuegos y gamificación (Erhel & Jamet, 2019; Giasiranis & Sofos, 2017; Hwang, Chiu, & Chen, 2015; Rodríguez-Ardura & Meseguer-Artola, 2017). La gamificación, que se define como el uso de mecanismos y elementos de diseño de los videojuegos en contextos educativos no recreativos con el propósito de incrementar el involucramiento y la experiencia del usuario (Domínguez et al., 2013), provee un mecanismo práctico para mejorar el proceso de aprendizaje, específicamente en aspectos motivacionales del estudiante (Chung, Shen, & Qiu, 2019).

Se ha encontrado evidencia de que cuando se logra el estado de flow en el contexto de gamificación, los estudiantes experimentan altos niveles de compromiso con las actividades escolares y motivación, de modo que el estado de flow ha sido considerado como un buen predictor del rendimiento escolar y de la calidad de la experiencia de aprendizaje (Joo, Oh, & Kim, 2015; Mesurado, 2010; Rijavec, Ljubin-Golub, Jurčec, & Olčar, 2017). En consecuencia, resulta imperativo utilizar un instrumento apropiado para evaluar el estado de flow en actividades de gamificación.

Aunque existen diversos instrumentos para evaluar el flow en el contexto de videojuegos y gamificación, tales como EGameFlow (Shu-Hui, Wann-Yih, & Dennison, 2018; Fu, Su, & Yu, 2009), Game Engagement Questionnaire (Brockmyer et al., 2009) o GameFlow Questionnaire (Kiili & Lainema, 2008), la Escala de Flow Disposicional-2 (DFS-2; Jackson & Eklund, 2002) ha sido reconocida como uno de los instrumentos más populares para la medición del flow en actividades generales, así como en estudios asociados con gamificación y videojuegos (Gutierrez, 2021; Hassan, Jylhä, Sjöblom, & Hamari, 2020; Marinho, Oliveira, Bittencourt, & Dermeval, 2019).

Además de la escala DFS-2, sus autores desarrollaron la escala FSS-2 (Flow State Scale-2), así como versiones cortas de estas escalas (S-DFS y S-FSS). Todas ellas se desarrollaron con base en modificaciones que los investigadores efectuaron a la versión original del instrumento (Jackson & Marsh, 1996), conocida como Flow State Scale (FSS), la cual fue desarrollada y validada en un contexto de actividades físicas y deportivas.

Tanto la escala DFS-2 como la escala FSS-2 evalúan las nueve dimensiones del flow propuestas por Csikszentmihalyi (1975). Para ambas escalas los ítems que las conforman son similares en su redacción, pero se presentan con diferentes modos de tiempo verbal para enfatizar distintos momentos de ocurrencia de la experiencia. La escala DFS-2 se utiliza para evaluar el flow disposicional, es decir, la tendencia general o disposición para experimentar el estado de flow dentro de un escenario señalado por el investigador o por el respondiente, en un marco de tiempo definido, en tanto que la escala FSS-2 evalúa el flow como un estado experimentado en un evento particular recién ocurrido (Jackson & Eklund, 2002). En esta investigación se propuso estudiar las propiedades psicométricas de la escala DFS-2 debido a su amplio reconocimiento y popularidad para la medición del flow en las experiencias de gamificación (Hamari & Koivisto, 2014; Marinho et al., 2019; Wang, Liu, & Khoo, 2009).

La escala DFS-2 fue investigada por sus desarrolladores en un estudio de validación cruzada utilizando una muestra de individuos que practicaban actividades físicas regularmente, con edades de 17 a 72 años (M = 26.3, DE = 10). Para el análisis, los autores propusieron dos modelos: (1) un modelo de medida, con nueve factores de primer orden (ver Figura 1), y (2) un modelo de segundo orden, con nueve factores de primer orden y un factor de segundo orden (ver Figura 2), que fueron evaluados por medio de un análisis factorial confirmatorio (AFC), utilizando CFI =

.90, NNFI = .90 y RMSEA = .08 como umbrales para los índices de ajuste de los modelos.

Como resultado de su investigación, Jackson y Eklund (2002) concluyeron que tanto el modelo de primer orden como el de segundo orden mostraron un ajuste aceptable (ver Tabla 1). Además, los valores de confiabilidad de cada dimensión del flow, medidos por medio del coeficiente alfa de Cronbach, se estimaron en un rango que va de .78 a .86, con un valor promedio de .82.

Las propiedades psicométricas de la escala DFS-2 han sido evaluadas en diversos estudios asociados con videojuegos y gamificación. Wang et al. (2009) estudiaron el desempeño de esta

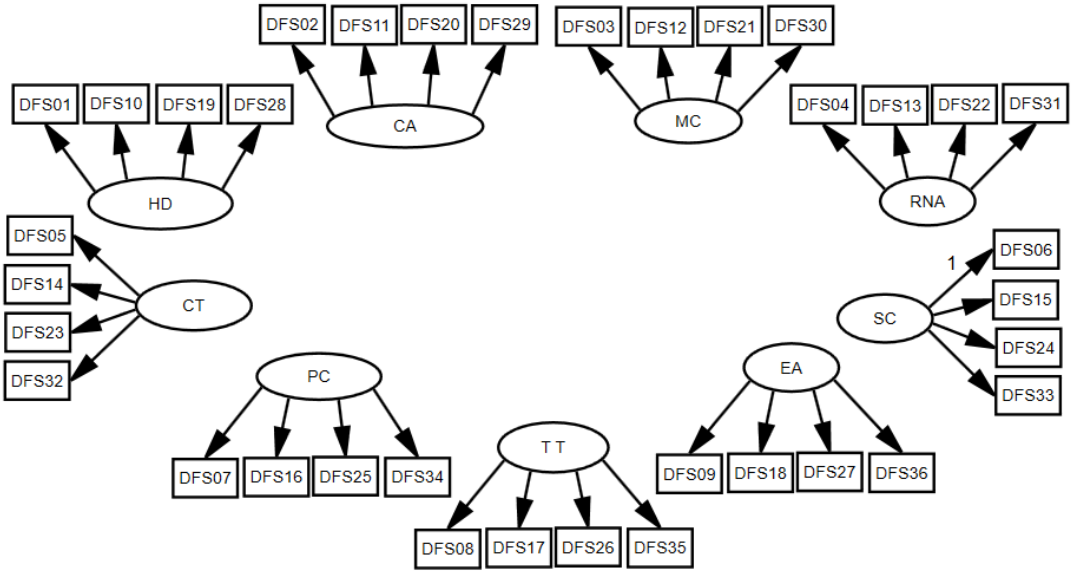


Figura 1
Estructura factorial de la DFS-2, modelo de primer orden (Jackson & Eklund, 2002).

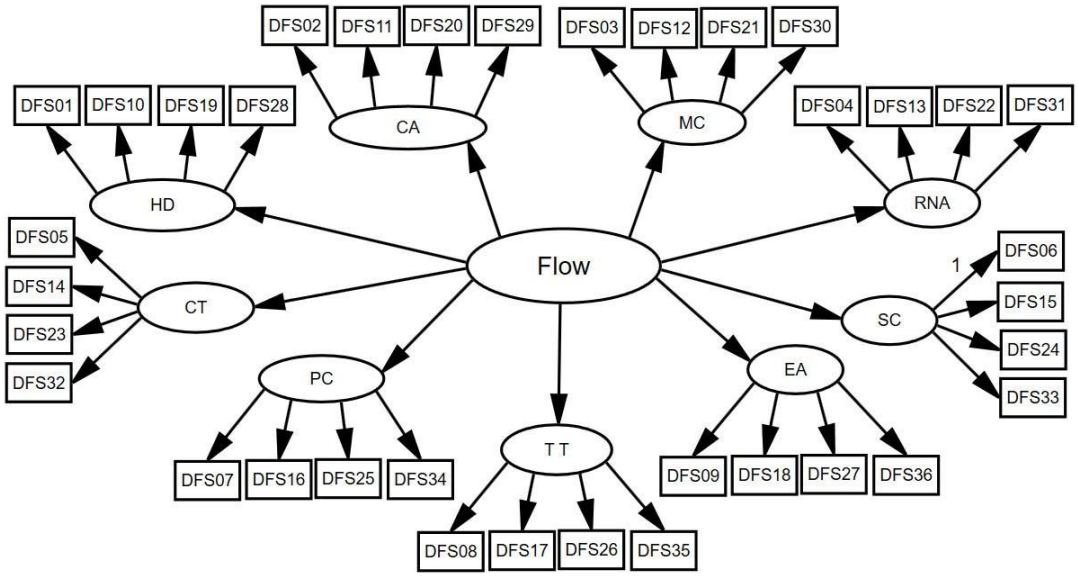


Figura 2
Estructura factorial de la DFS-2, modelo de segundo orden (Jackson & Eklund, 2002).

escala en el contexto de juegos en Internet, con base en una muestra de estudiantes de escuelas secundarias con edades entre 12 y 17 años (M = 13.2; DE = .80). Mediante un AFC con estimación robusta de máxima verosimilitud, y con un CFI = .95 y un NNFI = .95 como umbrales de ajuste, concluyeron que los modelos de Jackson y Eklund (2002) mostraron un buen ajuste, siendo mejor el ajuste para el modelo de segundo orden (ver Tabla 1).

Otros estudios que analizaron las propiedades psicométricas de la escala DFS-2 son los desarrollados por Procci, Singer, Levy y Bowers (2012) y por Hamari y Koivisto (2014). Los primeros estudiaron la escala utilizando una muestra de estudiantes universitarios que jugaban en computadoras, dispositivos móviles o consolas, con una edad promedio de 19.89 años (DE = 3.90), en tanto que los segundos utilizaron una muestra de usuarios de un servicio de actividad física por medio de gamificación, con una edad promedio de 29.5 años (DE = 27.5). En estos dos estudios los investigadores utilizaron los umbrales de ajuste CFI = .95 y RMSEA ≤ .06, siendo estos más estrictos que los utilizados por Jackson y Eklund (2002). De este modo, y con base en los resultados mostrados en la Tabla 1, concluyeron que la escala mostró un ajuste deficiente.

Aunque se han estudiado las propiedades psicométricas de la escala FSS traducida al español (García-Calvo, Jiménez-Castuera, Santos-Rosa-Ruano, Reina-Vaíllo, & Cervelló-Gimeno, 2008), y se han adaptado versiones cortas de la escala S-FSS al español (Calero & Injoque-Ricle, 2013), no se han estudiado las propiedades psicométricas de la escala DFS-2 para una versión en español en el contexto de los videojuegos. Esta investigación responde a esa necesidad planteándose como objetivo evaluar la adecuación de las propiedades psicométricas de la escala DFS-2 para su uso en la medición del flow en el contexto de los videojuegos.

Metodología

Participantes

Por medio de un esquema de muestreo no

Tabla 1
Estadísticos de ajuste en estudios de validación de la escala DFS-2, tal como se reportan en los trabajos originales.

Artículo	Modelo	N	χ^2	*df*	CFI	NNFI	RMSEA	90% IC
Jackson y Eklund (2002)	Primer orden	574	1427.2	558	.912	.901	.052	[.049, .055]
	Segundo orden	574	1606.5	585	.897	.889	.055	[.052, .058]
Wang et al. (2009)	Primer orden	1578	1925.5	558	.936	.927	.047	[.045, .049]
	Segundo orden	1578	1522.6	548	.954	.947	.040	[.038, .042]
Procci et al. (2012)	Primer orden	314	1351.9	558	.906	ND	.067	[.063, .072]
Hamari y Koivisto	Primer orden	200	1044.9	ND	.918	.907	.066	[.060, .073]
(2014)	Segundo orden	200	1136.6	ND	.906	.899	.069	[.063, .075]

Nota. Estadísticos de ajuste. χ^2: Chi-cuadrada, CFI: comparative fit index, NNFI: non-normed fit index, RMSEA: Steiger-Lind root mean square error of approximation. 90% IC: intervalo de confianza del 90% para el RMSEA. ND: no disponible.

probabilístico de conveniencia se recolectó la información de una muestra inicial de 326 estudiantes de nivel pregrado, de una universidad privada del estado de Nuevo León, México, durante el periodo de octubre a diciembre de 2019.

Con base en la identificación de datos atípicos de la muestra inicial se removieron 14 observaciones, de modo que la muestra final se conformó con N = 312 participantes, de los cuales el 40% eran mujeres y el 60% hombres, con un rango de edad de 16 a 34 años (M = 19.90; DE = 2.73). Los participantes manifestaron jugar videojuegos recreativos o educacionales de manera frecuente. El 34% de los participantes manifestó jugar cuatro o más veces por semana, en tanto que el 66% reportó jugar de una a tres veces por semana. La cantidad de horas promedio que los participantes dedican a jugar videojuegos, ya sea en computadoras, consolas o dispositivos móviles, fue de 4.50 horas por semana (DE = 5.33).

Instrumento

Los participantes contestaron la versión general de la DFS-2 traducida al idioma español a partir de la versión original en inglés. Esta traducción fue realizada por dos traductores profesionales y supervisada por el equipo de investigadores, contando con la autorización de los propietarios del instrumento (Jackson, Eklund, & Martin, 2020). Posteriormente se realizó la traducción en el sentido inverso, y se determinó que no se presentaban inconsistencias entre ambas traducciones. Dado que el contenido de la escala adaptada no fue alterado con respecto a su significado original, a juicio de los investigadores se determinó que la escala mostraba adecuada validez de contenido.

La escala DFS-2 evalúa la disposición para experimentar el flow con base en las nueve dimensiones propuestas por Csikszentmihalyi (1975), con cuatro ítems por dimensión, para sumar un total de 36 ítems. Estos se miden con una escala de tipo Likert de 5 puntos, donde 1 equivale a *fuertemente en desacuerdo* y 5 a *fuertemente de acuerdo*. Cada ítem se refiere a las experiencias y sentimientos que el participante percibió durante su participación en una actividad específica, ocurrida en un periodo de tiempo específico, ambos definidos por el investigador. Entre otras, algunas de las preguntas de la escala se refieren a cómo el participante percibe sus propias habilidades con relación al desafío propuesto en el contexto de la actividad analizada, su sentido de control y atención con respecto a lo que está haciendo, así como la satisfacción experimentada con la actividad. La distribución de los ítems del DFS-2 para cada una de las dimensiones del flow se muestra en la Figura 1.

Procedimiento

Los participantes respondieron libremente a una invitación para participar en el estudio, realizada a través de entrevistas personales, y a una convocatoria abierta por medios impresos y redes sociales al interior de la universidad. Una vez que otorgaron el consentimiento informado para participar en el estudio, cada participante contestó el instrumento de medición a través de un formulario en línea. En la recolección de datos no se recabó información sensible de los participantes.

La identificación y posterior remoción de datos atípicos se realizó por medio de las distancias de Mahalanobis, utilizando el criterio conservador $p < .001$ recomendado por Kline (2011), de modo que se obtuvo la muestra final de N = 312 participantes. De acuerdo con Kline (2011)

un tamaño de muestra de 200 casos corresponde al tamaño de muestra mediano reportado en estudios donde se utiliza el modelado con ecuaciones estructurales. Tabachnick y Fidell (2013) refieren que se requiere un tamaño de muestra mínimo de 300 casos para realizar un análisis factorial donde se presentan bajas comunalidades, pocos factores, y tres o cuatro indicadores por factor, en tanto que Costello y Osborne (2005) argumentan que una proporción de 10 observaciones por ítem es una regla ampliamente utilizada por los investigadores para determinar el tamaño de muestra. Basado en lo anterior, se concluyó que el tamaño de la muestra utilizada en este estudio era suficiente.

Para evaluar la validez de constructo se utilizó un AFC, así como también análisis factorial exploratorio (AFE). La validez convergente se evaluó por medio de los valores de la varianza media extraída de los constructos (AVE), así como con los coeficientes de confiabilidad alfa de Cronbach (α) y de confiabilidad compuesta (CR). La validez discriminante se evaluó empleando el criterio de Fornell-Larcker, además de la comparación de la AVE con respecto a la varianza compartida máxima (MSV) y la varianza compartida media (ASV). El análisis de datos se realizó por medio del software RStudio versión 1.3, utilizando el paquete lavaan (Rosseel, 2012) para el AFC y el paquete psych (Revelle, 2021) para el AFE. El nivel de significancia estadística se fijó en .05.

Resultado

Los resultados exploratorios, informados en la Tabla 2, muestran que las dimensiones de la escala DFS-2 con puntajes promedio más altos fueron metas claras, experiencia autotélica, retroalimentación no ambigua, sentido de control y balance habilidad-desafío, en tanto que los puntajes promedio más bajos se obtuvieron para las dimensiones concentración en la tarea, transformación del tiempo, mezcla de conciencia-acción y pérdida de conciencia.

AFC para los modelos originales de Jackson y Eklund

Antes de la evaluación del ajuste de los modelos de primer y segundo orden de Jackson y Eklund (2002) por medio de un AFC con estimación de parámetros por máxima verosimilitud, se analizó el supuesto de normalidad requerido para este método. Con base en las pruebas de Henze-Zirkler y de Royston, se encontró evidencia de violación del supuesto de distribución normal multivariada para los modelos analizados. En consecuencia, se utilizó el método de estimación de parámetros por máxima verosimilitud con errores estándar robustos y prueba Chi-cuadrada escalada de Satorra y Bentler (1994).

En modelos de ecuaciones estructurales, el

Tabla 2
Estadísticos descriptivos de las dimensiones de la escala DFS-2 con base en la muestra actual (N = 312).

Estadístico	MC	EA	RNA	SC	HD	CT	TT	CA	PC
Media	4.23	4.12	4.17	4.10	4.03	3.86	3.72	3.62	3.64
Desviación estándar	.63	.75	.65	.70	.67	.69	.94	.80	.99

Nota. Dimensiones: balance habilidad-desafío (HD), mezcla de conciencia-acción (CA), metas claras (MC), retroalimentación no ambigua (RNA), concentración en la tarea (CT), sentido de control (SC), pérdida de conciencia (PC), transformación del tiempo (TT), experiencia autotélica (EA).

estadístico más básico para probar el ajuste del modelo, bajo condiciones de normalidad multivariada de las variables observadas, es el estadístico Chi-cuadrada (χ^2). Cuando el estadístico χ^2 no es significativo, se tiene evidencia de que el modelo propuesto muestra un buen ajuste con la matriz de covarianza muestral (Tabachnick & Fidell, 2013). En esta investigación, el ajuste de los modelos se evaluó por medio del estadístico Chi-cuadrada escalada de Satorra-Bentler (χ^2_{SB}), además de los índices de ajuste CFI (comparative fit index), NNFI (non-normed fit index, también llamado TLI: Tucker-Lewis index), RMSEA (Steiger-Lind root mean square error of approximation) y SRMR (standardized root mean residual), utilizando como umbrales de ajuste aceptable los valores sugeridos por Hair, Black, Babin y Anderson (2014), Hu y Bentler (1999) y Tabachnick y Fidell (2013), los cuales se muestran en la Tabla 3.

Para los datos de la muestra actual, los modelos originales del DFS-2 de Jackson y Eklund (2002), de primer orden con 36 ítems agrupados en nueve constructos de cuatro ítems cada uno, así como de segundo orden jerárquico con nueve constructos de primer orden y un constructo de segundo orden, resultaron modelos sobreidentificados. Para ambos modelos, el estadístico Chi-cuadrada escalada de Satorra-Bentler resultó significativo ($p < .05$), lo que sugiere que el modelo no tiene un buen ajuste con los datos de la muestra.

Sin embargo, dado que el estadístico Chi-cuadrada tiende a mayores valores cuando se incrementa el tamaño de muestra o el número de variables observadas en el modelo (Hair et al., 2014), se procedió a obtener los valores de los índices de bondad de ajuste adicionales (ver Tabla 3). Al contrastar estos valores con los valores umbrales de ajuste aceptable sugeridos para un modelo de ecuaciones estructurales, se observó que los valores de CFI y NNFI eran mayores que .90, en tanto que se obtuvieron valores de RMSEA ≤ .06 y SRMR ≤ .08, tanto para el modelo de primer orden como para el modelo de segundo orden jerárquico. Así, la evidencia sugiere un ajuste aceptable de ambos modelos.

Con respecto a la evaluación de la validez convergente de la escala DFS-2, para todos los constructos se obtuvieron valores de los coeficientes α y CR mayores a .70 (ver Tabla 4) excepto para el constructo *sentido de control* (α = .68). Estos resultados sugieren adecuada consistencia interna de la escala (Hair et al., 2014; Moral de la Rubia, 2019; Taber, 2018). Así también, se observó que todos los parámetros de regresión resultaban significativos ($p < .001$) y con coeficientes estandarizados mayores que .50, a excepción del ítem DFS14 (ver Figura 3). Además, para todos los casos, excepto para el constructo *concentración con la tarea* (CT), los valores de AVE fueron mayores que .50, lo que sugiere aceptable validez convergente (Hair et al., 2014).

Al evaluar la validez discriminante, tal

Tabla 3
Estadísticos de ajuste para los modelos de la escala DFS-2 con base en la muestra actual (N = 312).

Especificación	χ^2_{SB}	*df*	CFI	NNFI	RMSEA	90% IC	SRMR
Primer orden	1041.50	558	.915	.904	.053	[.048, .057]	.056
Segundo orden	1185.40	585	.894	.886	.057	[.053, .062]	.071
Umbrales de ajuste			≥ .90	≥ .90	≤ .06		≤ .08

Nota. Estadísticos de ajuste: χ^2_{SB}: Chi-cuadrada escalada de Satorra-Bentler, CFI: comparative fit index, NNFI: non-normed fit index, RMSEA: Steiger-Lind root mean square error of approximation, 90% IC: intervalo de confianza del 90% para el RMSEA, SRMR: standardized root mean square residual.

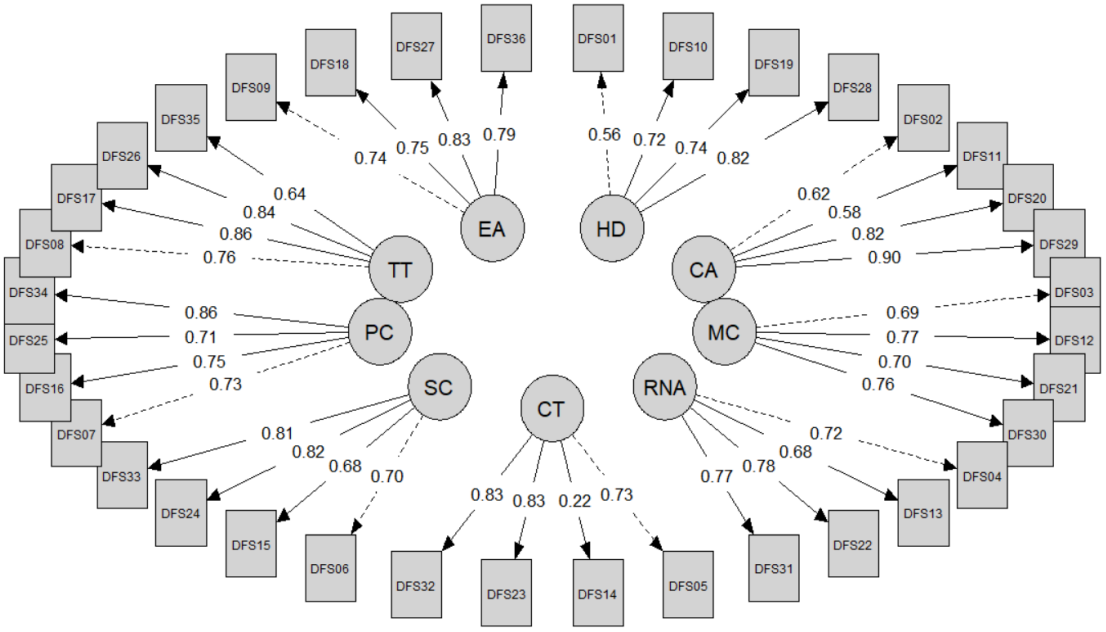


Figura 3
Estructura factorial del modelo de primer orden de la escala DFS-2, con coeficientes de regresión estandarizados.

como se muestra en la Tabla 4, se observaron correlaciones excesivamente altas para los pares de constructos RNA-MC, RNA-SC, MC-HD, RNA-HD, MC-SC, CT-SC y SC-HD. Así también, para los constructos HD, SC, RNA, CT y SC se encontraron valores de la AVE menores que los de su correspondiente MSV y ASV. Además, se observó que las raíces cuadradas de la AVE de los correspondientes constructos no cumplían en todos los casos con el criterio de Fornell-Larcker, que indica que estos valores deben ser mayores que las correlaciones con otros constructos (Fornell & Larcker, 1981). De esta forma se concluyó que para los datos de la muestra actual la escala DFS-2 no presentaba adecuada validez discriminante.

Reespecificación del modelo de primer orden de Jackson y Eklund

Al emplear la información de la muestra actual, los modelos originales de Jackson y Eklund (2002), de primer y segundo orden, mostraron un ajuste aceptable cuando se utilizan umbrales bajos para los índices de ajuste, además de aceptable validez convergente, sin embargo, mostraron falta de validez discriminante. Por ello, por medio de un AFE se procedió a identificar una estructura factorial que pudiera mostrar un mejor ajuste y validez discriminante.

Para el AFE se utilizó el método de extracción de factorización del eje principal y rotación oblicua promax, ya que se consideró que los constructos que conforman la escala se correlacionan. La medida de Kaiser-Meyer-Olkin (KMO = .94) mostró adecuación de la muestra. Así también, la prueba de esfericidad de Bartlett resultó significativa ($p < .001$), indicando que existe suficiente correlación entre las variables analizadas. Para determinar el número de factores a extraer se utilizó análisis paralelo y el diagrama de sedimentación. De esta forma se extrajeron cinco factores, que agruparon a 33 de los 36 ítems de la escala DFS-2, y que en conjunto explican el 53% de la varianza total, con valores para las cargas factoriales en el rango de .43 a .94 y comunalidades entre .46 y

Tabla 4
Correlaciones entre constructos para la escala DFS-2 con base en la muestra actual (N = 312).

Constructo	α	CR	AVE	MSV	ASV	HD	CA	MC	RNA	CT	SC	PC	TT	EA
HD	.81	.80	.52	.85	.56	**.72**								
CA	.81	.84	.55	.47	.29	.69	**.74**							
MC	.82	.82	.53	.97	.57	.92	.55	**.73**						
RNA	.83	.83	.55	.97	.58	.91	.56	.99	**.74**					
CT	.68	.75	.43	.80	.50	.76	.55	.80	.82	**.65**				
SC	.84	.84	.57	.94	.58	.87	.55	.95	.97	.89	**.75**			
PC	.84	.85	.58	.31	.22	.54	.36	.52	.50	.43	.56	**.76**		
TT	.85	.86	.60	.38	.16	.35	.41	.32	.29	.48	.35	.24	**.78**	
EA	.86	.86	.61	.61	.46	.77	.56	.74	.73	.78	.71	.48	.62	**.78**

Nota. Los valores en negritas representan las raíces cuadradas de la AVE de los constructos correspondientes. Constructos: balance habilidad-desafío (HD), mezcla de conciencia-acción (CA), metas claras (MC), retroalimentación no ambigua (RNA), concentración en la tarea (CT), sentido de control (SC), pérdida de conciencia (PC), transformación del tiempo (TT) y experiencia autotélica (EA). Índices: α: coeficiente alfa de Cronbach, CR: coeficiente de confiabilidad compuesta, AVE: varianza media extraída, MSV: varianza compartida máxima, ASV: varianza compartida media.

.80 (ver Tabla 5).

El primer factor (Maestría) agrupa la totalidad de los ítems de las dimensiones *sentido de control*, *retroalimentación no ambigua* y *metas claras*, así como tres ítems de la dimensión *concentración en la tarea* y dos ítems de la dimensión *balance habilidad-desafío*. A este factor se lo denominó *maestría en el juego*.

El segundo factor (TT) agrupa a todos los ítems de la dimensión *transformación del tiempo*. El tercer factor (PC) agrupa a todos los ítems de la dimensión *pérdida de conciencia*. El cuarto factor (CA) agrupa a todos los ítems de la dimensión *mezcla de conciencia-acción*, en tanto que el quinto factor (EA) agrupa a todos los ítems de la dimensión *experiencia autotélica*. Los ítems DFS01 y DFS14 mostraron comunalidades pequeñas, en tanto que el ítem DFS10 mostró una carga factorial pequeña. Por lo tanto, se consideró que estos ítems no aportaban significativamente a ninguno de los factores extraídos y no fueron incluidos en ninguna de las dimensiones identificadas.

Con los factores obtenidos mediante el AFE se propuso un modelo reespecificado de primer orden de la escala DFS-2 (ver Figura 4). Los resultados de un AFC mostraron que para este modelo el estadístico Chi-cuadrada escalada de Satorra-Bentler resultó significativo (χ^2_{SB} = 959.21, gl = 485, p < .001). Sin embargo, al observar los valores de los índices de ajuste se concluyó que el modelo muestra un ajuste aceptable (CFI = .910, NNFI = .902, SRMR = .057 y RMSEA = .056 con un 90% IC [.051, .061]). Además, se observó que todos los parámetros de regresión resultaron significativos (p < .001) y con coeficientes estandarizados mayores que .50 (ver Figura 4), lo que sugiere adecuada validez de constructo (Hair et al., 2014).

El modelo de primer orden reespecificado mostró adecuada validez convergente, ya que los valores de AVE para todos los constructos resultaron mayores que .50 (ver Tabla 6). Así también, se observó que, para todos los constructos, los valores de AVE resultaron mayores que los valores de MSV y ASV, excepto para el caso del

Tabla 5
Análisis factorial exploratorio con base en la muestra actual (N = 312).

		Factor					
Dimensión	**Ítem**	**Maestría**	**TT**	**PC**	**CA**	**EA**	**Comunalidades**
RNA	DFS22	**.85**	-.04	-.01	.11	-.11	.64
SC	DFS06	**.78**	.00	-.01	-.11	-.10	.53
MC	DFS03	**.78**	-.03	-.15	-.05	-.14	.54
SC	DFS33	**.76**	-.01	.03	.09	-.02	.65
MC	DFS30	**.75**	.00	.06	-.06	.01	.57
SC	DFS24	**.74**	.04	.04	.05	.06	.67
SC	DFS15	**.74**	.10	.22	-.08	-.25	.52
RNA	DFS04	**.70**	-.02	-.08	.01	-.06	.57
MC	DFS21	**.69**	-.02	.04	.03	.04	.51
RNA	DFS31	**.67**	-.03	.01	.00	.15	.59
RNA	DFS13	**.65**	-.02	.04	-.13	-.06	.53
CT	DFS32	**.64**	.08	-.02	.07	.18	.63
CT	DFS23	**.59**	.15	-.05	.04	.19	.59
CT	DFS05	**.57**	.13	-.13	-.08	.16	.46
MC	DFS12	**.51**	.00	.07	-.03	.09	.65
HD	DFS19	**.49**	-.08	.08	.19	.09	.50
HD	DFS28	**.43**	-.07	.04	.20	.18	.61
HD	DFS01	.27	-.12	.04	.07	.22	.31
TT	DFS17	.04	**.88**	.00	-.11	.05	.76
TT	DFS08	-.02	**.79**	-.02	.00	-.03	.58
TT	DFS26	.04	**.78**	-.08	-.09	.18	.69
TT	DFS35	-.09	**.55**	.01	.18	.09	.46
PC	DFS34	-.02	-.03	**.87**	-.06	.14	.74
PC	DFS16	.10	.04	**.76**	-.04	-.10	.60
PC	DFS07	-.07	-.07	**.76**	-.05	.12	.53
PC	DFS25	.07	-.05	**.61**	.01	.11	.48
CA	DFS29	.02	-.01	-.08	**.94**	.10	.80
CA	DFS20	-.05	-.05	.00	**.91**	.01	.69
CA	DFS02	.24	-.05	-.14	**.52**	-.01	.41
CA	DFS11	-.03	.03	.00	**.43**	-.09	.51
CT	DFS14	-.05	.11	.27	.30	-.16	.26
EA	DFS27	.01	.04	.09	.03	**.76**	.73
EA	DFS36	.07	.11	.05	-.01	**.63**	.59
EA	DFS09	.06	.08	.01	-.08	**.59**	.63
EA	DFS18	.16	.20	.05	-.02	**.50**	.55
HD	DFS10	.32	-.08	-.02	.06	.22	.60
Varianza explicada		25%	7%	7%	7%	7%	

Nota. Las cargas factoriales significativas se muestran en negritas. Dimensiones del flow: balance habilidad-desafío (HD), mezcla de conciencia-acción (CA), metas claras (MC), retroalimentación no ambigua (RNA), concentración en la tarea (CT), sentido de control (SC), pérdida de conciencia (PC), transformación del tiempo (TT) y experiencia autotélica (EA). Maestría: maestría en el juego.

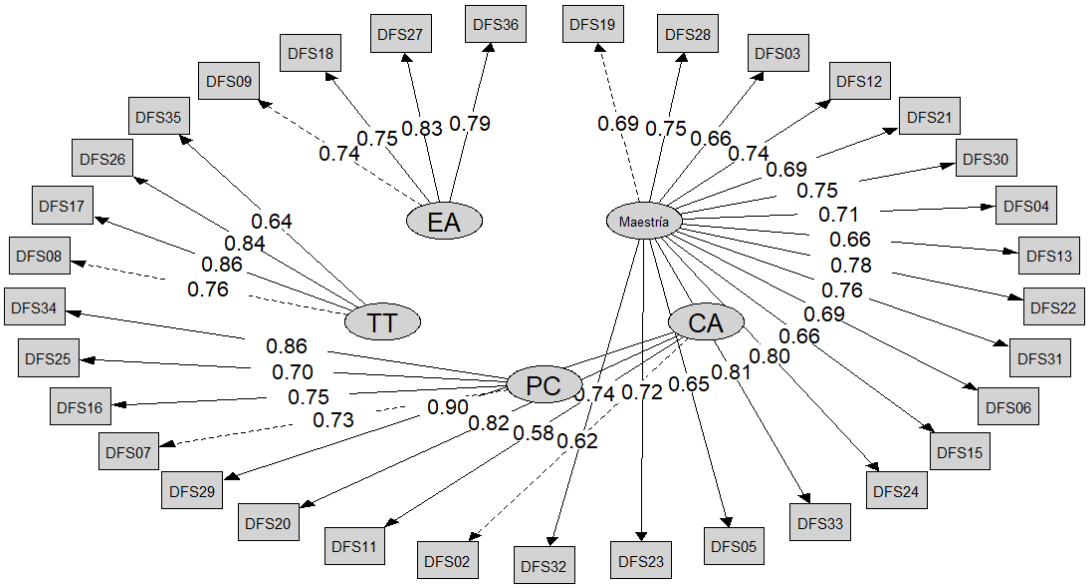


Figura 4
Estructura factorial del modelo reespecificado para la escala DFS-2 con coeficientes de regresión estandarizados.

constructo maestría en el juego, donde el valor de AVE resultó menor que el de MSV. Además, las correlaciones entre constructos no resultaron demasiado altas, en tanto el criterio de Fornell-Larcker se cumplió satisfactoriamente para todos los constructos, excepto en el caso de maestría, donde se observó un valor del coeficiente de correlación entre los constructos experiencia autotélica y maestría en el juego ligeramente mayor que la raíz cuadrada del AVE correspondiente. De esta forma se concluyó que el modelo de primer orden reespecificado de la escala DFS-2 mostraba adecuada validez discriminante.

Discusión

En este estudio se evaluaron las propiedades psicométricas de la escala DFS-2 de Jackson y Eklund (2002) en el contexto de videojuegos recreativos con base en una muestra de estudiantes

Tabla 6
Correlaciones entre constructos para el modelo reespecificado de la escala DFS-2.

Constructo	α	CR	AVE	MSV	ASV	Maestría	CA	PC	TT	EA
Maestría	.95	.95	.52	.59	.34	**.72**				
CA	.81	.83	.55	.35	.24	.59	**.74**			
PC	.84	.85	.58	.29	.17	.54	.36	**.76**		
TT	.85	.86	.60	.38	.19	.37	.41	.24	**.78**	
EA	.86	.86	.61	.59	.38	.77	.56	.48	.62	**.78**

Nota. Los valores en negritas representan las raíces cuadradas de AVE de los constructos correspondientes. Constructos: maestría en el juego (Maestría), mezcla de conciencia-acción (CA), pérdida de conciencia (PC), transformación del tiempo (TT) y experiencia autotélica (EA). Índices: α: coeficiente alfa de Cronbach, CR: coeficiente de confiabilidad compuesta, AVE: varianza media extraída, MSV: varianza compartida máxima, ASV: varianza compartida media.

de una universidad mexicana. Para los modelos originales de primer y segundo orden de la escala, los valores de los índices de ajuste aproximados obtenidos sugirieron un ajuste aceptable de ambos modelos con respecto a los datos de la muestra, de forma similar a los resultados obtenidos en el estudio de validación cruzada realizado por los autores de la escala. Al comparar los índices de ajuste para ambos modelos, se observó un mejor ajuste para el modelo de primer orden.

Los valores de los índices de ajuste para los modelos originales de primer y segundo orden de la escala DFS-2 encontrados en este estudio son similares a los informados en los trabajos de Hamari y Koivisto (2014), Jackson y Eklund (2002) y Procci et al. (2012), y ligeramente inferiores a los informados por Wang et al. (2009). Sin embargo, si se establecen umbrales más estrictos para estos índices, tal como lo proponen Hamari y Koivisto (2014) y Procci et al. (2012), ambos modelos muestran un ajuste pobre.

Por otra parte, aunque la escala mostró adecuada validez convergente, se encontró evidencia de inadecuada validez discriminante. Este resultado contrasta con los presentados en estudios similares, pero realizados en contextos ajenos a los videojuegos, donde se ha encontrado evidencia de adecuada validez convergente y validez discriminante para la escala DFS-2 (Bittencourt et al., 2021; Riva et al., 2017). No obstante, dado que Swann, Crust y Vella (2017) han criticado la validez discriminante de la escala FSS-2, y esta a su vez tiene una estructura similar a la escala DFS-2, cabe la posibilidad de que para ciertos contextos asociados con videojuegos la validez discriminante de la escala DFS-2 no se satisfaga.

La inadecuada validez discriminante sugiere que algunas dimensiones de la escala no difieren suficientemente de otras como para considerarse distintas (Hair et al., 2014). Este fenómeno podría atribuirse a que en el modelo de la escala DFS-2, Jackson y Eklund (2002) no tomaron en cuenta las relaciones entre las dimensiones del flow (Hamari & Koivisto, 2014). Una conceptualización del flow en el contexto de los videojuegos que considera estas relaciones es la que plantean Hamari y Koivisto (2014), así como Kiili et al. (2012). Estos autores postulan que el flow puede segmentarse en antecedentes y consecuencias.

De acuerdo con Kiili et al. (2012) las dimensiones del flow pueden agruparse en antecedentes, estado de flow y consecuencias. De acuerdo a este enfoque, los antecedentes del flow, también llamados precondiciones del flow (Khoshnoud et al., 2020) agrupan las dimensiones *metas claras*, *sentido de control*, *balance habilidad-desafío*, *retroalimentación no ambigua* y *jugabilidad*. Esta última reemplaza a la dimensión *mezcla de conciencia-acción*, y se utiliza en el contexto del diseño y análisis de videojuegos para "describir la calidad del juego en términos de sus reglas, mecanismos, metas y diseño" (Gonzalez, 2013, p. 92). De igual forma, Hamari y Koivisto (2014) agrupan a las mismas dimensiones, a excepción de la dimensión *jugabilidad* y añadiendo la dimensión *experiencia autotélica*, en un constructo llamado *maestría*, definida como la capacidad del jugador para adquirir las habilidades para desempeñarse en el juego. Así, el constructo maestría representa el antecedente del flow.

En este estudio las intercorrelaciones más altas se encontraron entre las dimensiones propuestas como antecedentes del flow por Hamari y Koivisto (2014) y Kiili et al. (2012), lo cual motivó una reespecificación del modelo original de primer orden de la escala DFS-2 por medio de un AFE y un AFC. De esta forma se identificó un modelo reespecificado de primer orden con cinco dimensiones: (1) transformación del tiempo, (2) pérdida de conciencia, (3) mezcla de conciencia-acción, (4) experiencia autotélica y (5) maes-

tría en el juego.

De manera similar que las propuestas de Procci et al. (2012) y Hamari y Koivisto (2014), la dimensión *maestría en el juego*, que agrupa a los ítems que parecen representar la capacidad del jugador para adquirir las habilidades para un buen desempeño en el juego, representa el antecedente del flow; en tanto que las otras cuatro dimensiones del modelo reespecificado de la escala DFS-2 parecen representar las consecuencias del flow. Este modelo mostró propiedades psicométricas adecuadas, ya que se obtuvo un ajuste aceptable con respecto a los datos de la muestra, similar al encontrado para el modelo original de Jackson y Eklund (2002), así como adecuada validez convergente y validez discriminante.

Los hallazgos de esta investigación, acerca de la segmentación del flow en antecedentes y consecuencias, están en línea con las propuestas de Hamari y Koivisto (2014) y Kiili et al. (2012), y sugieren la necesidad de investigaciones futuras encaminadas a adaptar la escala DFS-2 al idioma español en el contexto de videojuegos y gamificación, con el objetivo de encontrar una estructura factorial que posea propiedades psicométricas adecuadas para medir el flow y las dimensiones que lo conforman. Cabe mencionar que para este estudio no fue posible obtener una segunda muestra para la validación del modelo factorial reespecificado con cinco dimensiones identificado con base en el AFC. De este modo, se propone para futuras investigaciones validar este modelo con muestras adicionales.

Adicionalmente, ya que la escala DFS-2 fue creada como una variante de la escala FSS, desarrollada originalmente para evaluar el estado de flow en actividades físicas y deportivas, se recomienda a los investigadores analizar las propiedades psicométricas de otras escalas, tales como EGameFlow, Game Engagement Questionnaire o GameFlow, de modo que se disponga de opciones adicionales para la medición del flow en el contexto de los videojuegos y la gamificación en usuarios de habla hispana.

Finalmente, aunque los hallazgos de esta investigación se derivan de una muestra no probabilística, lo cual no permite la generalización de las conclusiones, los resultados obtenidos sugieren que para la evaluación del estado de flow en videojuegos o actividades de gamificación, la escala DFS-2 original muestra adecuada validez de constructo y validez convergente, sin embargo, no muestra adecuada validez discriminante. De acuerdo con Farrel (2010) una solución a la falta de validez discriminante consistiría en combinar las dimensiones o constructos en una sola medida global, en lugar de realizar un análisis dimensión por dimensión. Así, en el contexto de videojuegos y gamificación, la escala DFS-2 original puede ser útil para los investigadores que deseen evaluar el flow como un constructo global, sin necesidad de discriminar en sus nueve dimensiones.

Referencias

Bittencourt, I. I., Freires, L., Lu, Y., Challco, G. C., Fernandes, S., Coelho, J., ... & Isotani, S. (2021). Validation and psychometric properties of the Brazilian-Portuguese dispositional Flow Scale 2 (DFS-BR). *PLoS ONE 16*(7), e0253044. doi: 10.1371/journal.pone.0253044

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology, 45*(4), 624-634. doi: 10.1016/j.jesp.2009.02.016

Calero, A., & Injoque-Ricle, I. (2013). Propiedades psicométricas del Inventario Breve de Experiencias Óptimas (Flow). *Revista Evaluar, 13*(1), 40-55. doi:

10.35670/1667-4545.v13.n1.6796

Chung, C. H., Shen, C., & Qiu, Y. Z. (2019). Students' acceptance of gamification in Higher Education. *International Journal of Game-Based Learning, 9*(2), 1-19. doi: 10.4018/IJGBL.2019040101

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*, Article 7. doi: 10.7275/jyj1-4868

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety.* San Francisco, CA: Jossey-Bass.

Csikszentmihalyi, M. (2014). Play and intrinsic rewards. En M. Csikszentmihalyi (Ed.), *Flow and the foundations of positive psychology. The collected works of Mihaly Csikszentmihalyi* (pp. 135-154). Nueva York, NY: Springer. doi: 10.1007/978-94-017-9088-8

Domínguez, A., Saenz de Navarrete, J., de Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J. J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education, 63*, 380-392. doi: 10.1016/j.compedu.2012.12.020

Erhel, S., & Jamet, E. (2019). Improving instructions in educational computer games: Exploring the relations between goal specificity, flow experience and learning outcomes. *Computers in Human Behavior, 91*, 106-114. doi: 10.1016/j.chb.2018.09.020

Farrel, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research, 63*(3), 324-327. doi: 10.1016/j.jbusres.2009.05.003

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. doi: 10.1177/002224378101800104

Fu, F. -L., Su, R. -C., & Yu, S. -C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education, 52*(1), 101-112. doi: 10.1016/j.compedu.2008.07.004

García-Calvo, T., Jiménez-Castuera, R., Santos-Rosa-Ruano, F. J., Reina-Vaíllo, R., & Cervelló-Gimeno, E. (2008). Psychometric properties of the Spanish version of the Flow State Scale. *The Spanish Journal of Psychology, 11*(2), 660-669. doi: 10.1017/s1138741600004662

Giasiranis, S., & Sofos, L. (2017). Flow experience and educational effectiveness of teaching informatics using AR. *Educational Technology & Society, 20*(4), 78-88. Recuperado de http://www.jstor.org

Gonzalez, C. (Ed.). (2013). *Student usability in educational software and games*: *Improving experiences*. Hershey, PA: IGI Global. doi: 10.4018/978-1-4666-1987-6

Gutierrez, J. P. (2021). Do game transfer phenomena lead to flow? An investigation of in-game and out-game immersion among MOBA gamers. *Computers in Human Behavior Reports, 3*, Article 100079. doi: 10.1016/j.chbr.2021.100079

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7ª ed.). Harlow, Reino Unido: Pearson Education Limited.

Hamari, J., & Koivisto, J. (2014). Measuring flow in gamification: Dispositional Flow Scale-2. *Computers in Human Behavior, 40*, 133-143. doi: 10.1016/j.chb.2014.07.048

Hassan, L., Jylhä, H., Sjöblom, M., & Hamari, J. (2020). Flow in VR: A study on the relationships between preconditions, experience and continued use. *Proceedings of the 53rd Hawaii International Conference on System Sciences* (pp. 1196-1205). doi: 10.24251/HICSS.2020.149

Hu, L. -T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Hwang, G. -J., Chiu, L. -Y., & Chen, C. -H. (2015). A contextual game-based learning approach to improving students´ inquiry-based learning performance in social studies courses. *Computers & Education, 81*, 13-25. doi: 10.1016/j.compedu.2014.09.006

Jackson, S. A., & Eklund, R. C. (2002). Assessing flow in

physical activity: The Flow State Scale-2 and Dispositional Flow Scale-2. *Journal of Sport & Exercise Psychology, 24*(2), 133-150. doi: 10.1123/jsep.24.2.133

Jackson, S. A., Eklund R. C., & Martin, A. J. (23 de junio de 2020). *Flow Scales*. Mind Garden. Recuperado de https://www.mindgarden.com/100-flow-scales

Jackson, S. A., & Marsh, H. W. (1996). Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of Sport & Exercise Psychology, 18*(1), 17-35. doi: 10.1123/jsep.18.1.17

Joo, Y. J., Oh, E., & Kim, S. M. (2015). Motivation, instructional design, flow, and academic achievement at a Korean online university: A structural equation modeling study. *Journal of Computing in Higher Education, 27*(1), 28-46. doi: 10.1007/s12528-015-9090-9

Khoshnoud, S., Alvarez-Igarzábal, F., & Wittmann, M. (2020). Peripheral-physiological and neural correlates of the flow experience while playing video games: A comprehensive review. *PeerJ, 8*, e10520. doi: 10.7717/peerj.10520

Kiili, K., de Freitas, S., Arnab, S., & Lainema, T. (2012). The design principles for flow experience in educational games. *Procedia Computer Science, 15,* 78-91. doi: 10.1016/j.procs.2012.10.060

Kiili, K., & Lainema, T. (2008). Foundation for measuring engagement in educational games. *Journal of Interactive Learning Research, 19*(3), 469-488. Recuperado de https://www.learntechlib.org

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3ª ed.). New York, NY: The Guilford Press.

Marinho, A., Oliveira, W., Bittencourt, I. I., & Dermeval, D. (2019). Does gamification improve flow experience in classroom? An analysis of gamer types in collaborative and competitive settings. *Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 27*(2), 40-68. doi: 10.5753/RBIE.2019.27.02.40

Mesurado, B. (2010). La experiencia de flow o experiencia óptima en el ámbito educativo. *Revista Latinoamericana de Psicología, 42*(2), 183-192. Recuperado de http://revistalatinoamericanadepsicologia.konradlorenz.edu.co

Montes-González, J. A., Ochoa-Angrino, S., Baldeón-Padilla, D. S., & Bonilla-Sáenz, M. (2018). Videojuegos educativos y pensamiento científico: Análisis a partir de los componentes cognitivos, metacognitivos y motivacionales. *Educación y Educadores, 21*(3), 388-408. doi: 10.5294/edu.2018.21.3.2

Moral de la Rubia, J. (2019). Revisión de los criterios para validez convergente estimada a través de la Varianza Media Extraída. *Psychologia, 13*(2), 25-41. doi: 10.21500/19002386.4119

Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.

Procci, K., Singer, A. R., Levy, K. R., & Bowers, C. (2012). Measuring the flow experience of gamers: An evaluation of the DFS-2. *Computers in Human Behavior, 28*(6), 2306-2312. doi: 10.1016/j.chb.2012.06.039

Revelle, W. (2021). psych: Procedures for psychological, psychometric, and personality research (R package 2.1.9). [Software de cómputo]. Recuperado de https://CRAN.R-project.org/package=psych

Rijavec, M., Ljubin-Golub, T., Jurčec, L., & Olčar, D. (2017). Working part-time during studies: The role of flow in students' well-being and academic achievement. *Croatian Journal of Education, 19*. doi: 10.15516/cje.v19i0.2724

Riva, E. F. M., Riva, G., Talò, C., Boffi, M., Rainisio, N., Pola, L., ... & Inghilleri, P. (2017). Measuring dispositional flow: Validity and reliability of the Dispositional Flow State Scale 2, Italian version. *PLoS ONE, 12*(9), e0182201. doi: 10.1371/journal.pone.0182201

Rodríguez-Ardura, I., & Meseguer-Artola, A. (2017). Flow in e-learning: What drives it and why it matters. *British Journal of Educational Technology, 48*(4), 899-915. doi: 10.1111/bjet.12480

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. doi: 10.18637/jss.v048.i02

Satorra, A., & Bentler, P. M. (1994). Correction to test statistics and standard errors in covariance structure analysis. En A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Shu-Hui, C., Wann-Yih, W., & Dennison, J. (2018). Validation of EGameFlow: A self-report scale for measuring user experience in video game play. *Computers in Entertainment, 16*(3), 1-15. doi: 10.1145/3238249

Stavrou, N. A., & Zervas, Y. (2004). Confirmatory factor analysis of the flow state scale in sports. *International Journal of Sport and Exercise Psychology, 2*(2), 161-181. doi: 10.1080/1612197X.2004.9671739

Swann, C., Crust, L., & Vella, S. A. (2017). New directions in the psychology of optimal performance in sport: Flow and clutch states. *Current Opinion in Psychology, 16*, 48-53. doi: 10.1016/j.copsyc.2017.03.032

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6ª ed.). Upper Saddle River, NJ: Pearson.

Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273-1296. doi: 10.1007/s11165-016-9602-2

Wang, C. K. J., Liu, W. C., & Khoo, A. (2009). The psychometric properties of Dispositional Flow Scale-2 in internet gaming. *Current Psychology, 28*(3), 194-201. doi: 10.1007/s12144-009-9058-x

2021, Vol. 21, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

Cuestionario de Acciones Creativas: Propiedades psicométricas de la versión abreviada (CAC42)

Creative Actions Questionnaire: Psychometric Properties of the Abbreviated Version (CAC42)

Romina Cecilia Elisondo * [1, 2] , Danilo Silvio Donolo [1]

1 - Universidad Nacional de Río Cuarto Ruta 36 Km. 601, Río Cuarto.
2 - Consejo Nacional de Investigaciones Científicas y Técnicas CONICET, Argentina.

Recibido: 25/02/2021 **Revisado:** 03/05/2021 **Aceptado:** 07/05/2021

Resumen

El objetivo es construir una versión abreviada del Cuestionario de Acciones Creativas (CAC) y analizar sus propiedades psicométricas. El CAC incluye escalas que evalúan la frecuencia de participación en acciones creativas en siete dominios: *literatura*; *artes plásticas y artesanías*; *ciencia y tecnología*; *artes escénicas*; *música*; *participación social* y *creatividad cotidiana*. La muestra se conformó con 1509 personas mayores de 18 años que residen en diferentes provincias argentinas. Los instrumentos de recolección de datos fueron: CAC, Inventario Biográfico de Comportamientos Creativos (BICB), cuestionario sociodemográfico y de ocio. El análisis factorial confirmatorio demostró un adecuado ajuste del modelo. Los resultados indican diferencias significativas según participación en actividades de ocio. Observamos correlaciones significativas y moderadas entre la versión abreviada del CAC y el BICB. Los resultados señalan que la versión abreviada del CAC cumple con estándares técnicos y constituye un aporte para la valoración de acciones creativas en diferentes dominios.

Palabras clave: *acciones creativas, evaluación de la creatividad, creatividad cotidiana, participación social*

Abstract

The objective is to build an abbreviated version of the Creative Actions Questionnaire (CAC) and analyze its psychometric properties. The CAC includes scales that evaluate the frequency of participation in creative actions in seven domains: *literature*; *plastic arts and crafts*; *science and technology*; *performing arts*; *music*; *social participation* and *daily creativity*. The sample was made up of 1509 people over 18 years of age who reside in different Argentine provinces. The data collection instruments were: CAC, Biographical Inventory of Creative Behaviors (BICB), and sociodemographic and leisure questionnaire. The confirmatory factor analysis showed an adequate fit of the model. The results indicate significant differences according to participation in leisure activities. We observe significant and moderate correlations between the abbreviated version of the CAC and the BICB. The results show that the abbreviated version of the CAC complies with technical standards and constitutes a contribution for the evaluation of creative actions in different domains.

Keywords: *creative actions, evaluation of creativity, everyday creativity, social participation*

***Correspondencia a:** Romina Cecilia Elisondo. Tel: 0358-4388581. E-mail: relisondo@gmail.com

Cómo citar este artículo: Elisondo, R. C., & Donolo, D. S. (2021). Cuestionario de Acciones Creativas: Propiedades psicométricas de la versión abreviada (CAC42). *Revista Evaluar, 21*(3), 81-94. Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Nota de autor: Este trabajo fue aprobado como proyecto y subsidiado por el Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina CONICT (PIP 11220130100474) para el período comprendido entre los años 2015 y 2020.

Participaron en la edición de este artículo: Florencia Ruiz, Alicia Molinari, Mónica Serppe, Andrea Suárez, Juan Balverdi, Benjamín Casanova, Ricardo Hernández.

Introducción

De acuerdo con especialistas (Diedrich et al. 2018; Glăveanu et al., 2019; Silvia, Cotter, & Christensen, 2017) aceptamos que las personas generamos ideas, productos y acciones creativas en diferentes situaciones de nuestra vida cotidiana. Richards (2010) define a la creatividad cotidiana como la originalidad humana en el trabajo, el ocio y las diversas actividades de la vida diaria. Según la autora, la creatividad cotidiana es indispensable para la supervivencia humana y se basa en el compromiso y la participación activa de los sujetos.

En el campo de investigación de la creatividad se han construido diversas estrategias para valorar las acciones cotidianas de los sujetos. Entre las técnicas más utilizadas se encuentran los cuestionarios que refieren a comportamientos creativos, instrumentos que permiten valorar la frecuencia con que las personas desarrollan determinadas acciones creativas (Aranguren & Irrazabal, 2012; Carson, Peterson, & Higgins, 2005; Diedrich et al., 2018; Paek & Runco, 2017). También se han desarrollado procedimientos como la evaluación momentánea ecológica que intenta captar los desempeños diarios de las personas en sus contextos naturales. En estos diseños, se realizan observaciones repetidas de los fenómenos para analizar el dinamismo de las acciones creativas (Silvia et al., 2017). Los estudios sobre capacidades creativas y personalidad también permiten comprender la complejidad de las manifestaciones diarias de la creatividad. Se han observado interesantes correlaciones entre creatividad, apertura a la experiencia y extraversión (Elisondo, Donolo, & Corbalan-Berná, 2009; Garcês et al., 2015), es decir, las personas más creativas se caracterizan por ser sociables y abiertas a desarrollar experiencias en contextos heterogéneos. Asimismo, se han desarrollado estudios cualitativos que procuran analizar acciones creativas a partir de significaciones y valoraciones de grupos diversos. Los estudios muestran que las personas desarrollan procesos de creatividad cotidiana a partir de diferentes acciones creativas, generalmente en interacción con otras personas y en contextos particulares (hogar, ONGs, instituciones educativas, etc.). Los sujetos reconocen impactos positivos de la creatividad cotidiana en la calidad de vida, el bienestar y salud en sentido amplio (Adams-Price & Steinman, 2007; Elisondo & Vargas, 2019; Gandolfo & Grace, 2010; Genoe & Liechty, 2017; McCabe & de Waal-Malefyt, 2015; Modrzejewska-Świgulska, 2018).

El objetivo del estudio es construir el CAC42, versión abreviada del Cuestionario de Acciones Creativas (Elisondo & Donolo, 2016), y analizar sus propiedades psicométricas. En el estudio examinamos la estructura interna del cuestionario y presentamos evidencias de validez convergente, para ello utilizamos el Inventario Biográfico de Comportamientos Creativos (Batey, 2007). Considerando investigaciones previas que indican diferencias en acciones creativas según sexo, escolaridad y participación en actividades de ocio (Aranguren & Irrazabal, 2012; Diedrich et al., 2018; Elisondo & Donolo, 2016; Kaufman, 2006; Paek & Runco, 2017), analizamos en el presente estudio diferencias en los dominios del CAC42 según dichas variables. En general, los estudios indican puntajes superiores de las mujeres en artesanías y áreas artísticas, y diferencias a favor de los hombres en dominios tecnológicos (Diedrich et al., 2018; Paek & Runco, 2017). Las investigaciones también mostraron que los participantes con mayor nivel educativo obtienen puntajes superiores en acciones creativas (Aranguren & Irrazabal, 2012; Elisondo & Donolo, 2016). Asimismo, estudios previos indican que las personas que participan activamente en propuestas de ocio logran puntajes superiores en las escalas

de acciones creativas (Aranguren & Irrazabal, 2012; Hegarty, 2009; Wolfradt & Pretz, 2001).

Subrayamos la importancia de los estudios de la creatividad cotidiana, considerando las numerosas evidencias que indican relaciones entre salud, bienestar y desarrollo de acciones creativas en contextos laborales y de ocio (Benedek, Bruckdorfer, & Jauk, 2019; Corner & Silvia, 2015; Conner, DeYoung, & Silvia, 2018; Richards, 2010). Destacamos la relevancia del estudio en tanto contribución al campo de la evaluación de la creatividad en Argentina, aportando una herramienta que puede ser de utilidad para el diagnóstico y la investigación de perfiles creativos en contextos educativos, laborales y de tiempo libre. En el CAC42 también es necesario subrayar el enfoque sociocultural adoptado y el énfasis puesto en las interacciones sociales. Los ítems del instrumento se elaboraron considerando la importancia de la participación en grupos y el reconocimiento social de otras personas. En suma, el CAC42 ha sido construido conforme a enfoques actuales del estudio de la creatividad e intenta valorar acciones creativas en dominios diversos, no solo considerando las áreas artísticas y científicas tradicionalmente valoradas en cuestionarios de este tipo.

Evaluación de la creatividad: dominio general vs. dominios específicos

La evaluación de la creatividad es un campo por demás controvertido, en tanto refiere a un constructo difícil de definir y operacionalizar. Si bien se han realizado significativos avances, se observan inconsistencias en los resultados, dificultades técnicas de los instrumentos y falta de acuerdo en las definiciones teóricas del constructo (Acar & Runco, 2019; Barbot, Hass, & Reiter-Palmon, 2019). En cuanto a las definiciones, el debate acerca de la creatividad como dominio general o específico tiene importantes derivaciones teóricas, metodológicas y prácticas. Las posturas que sostienen que la creatividad es un dominio general proponen instrumentos unidimensionales para la valoración de capacidades creativas y pensamientos divergentes. Por ejemplo, el test CREA se sustenta en concepciones de la inteligencia creativa como capacidad general de formular interrogantes (Elisondo & Donolo, 2018). En cambio, quienes sostienen que la creatividad adquiere características particulares según diferentes dominios apoyan la construcción de instrumentos que valoran desempeños creativos en diferentes áreas (Diedrich et al., 2018; Kaufman, 2012; Paek & Runco, 2017). Según Baer (2012), un creciente cuerpo de evidencia de investigación sugiere que la creatividad es específica según cada dominio y que las habilidades generales contribuyen poco al desempeño creativo. Según el investigador, estos avances tienen importantes implicancias en las formas de entender, evaluar y promover la creatividad. El instrumento que se propone en el presente artículo se apoya en concepciones que definen a la creatividad como dominio específico. Sin embargo, no se desconoce la incidencia de conocimientos y habilidades generales en los procesos creativos. En este sentido, los modelos actuales incluyen componentes específicos y generales. Sternberg (2012) señala que la creatividad no es de dominio general ni específico, sino que conjuga ambos elementos. El autor señala que la especificidad de dominio de la creatividad depende de una base general de conocimientos y de decisiones respecto del uso de esta base. Kaufman (2012) propone un modelo jerárquico que incluye requisitos iniciales para la creatividad (inteligencia, motivación, etc.), áreas temáticas generales (por ejemplo, escritura, ciencia), dominios (poesía o ficción) y micro-dominios (por ejemplo, *haikus* o verso libre). Según el investigador, cada

área temática puede tener su propio perfil de personalidad y de patrones cognitivos.

En suma, los debates sobre la generalidad o especificidad de la creatividad, lejos de estar saldados, cada vez se complejizan más, especialmente por los aportes de tecnologías más sofisticadas para el estudio de los procesos creativos. En esta línea, los estudios neuropsicológicos muestran interesantes resultados con respecto a los procesos cognitivos generales y específicos (Benedek, Christensen, Fink, & Beaty, 2019), abriendo numerosas posibilidades para investigaciones futuras.

Escalas de comportamientos creativos

El estudio de las manifestaciones cotidianas de la creatividad es un área de notorio interés en el campo de investigación de los procesos creativos. Los investigadores han construido diferentes estrategias y procedimientos con el propósito de indagar comportamientos y logros creativos de las personas en diferentes dominios de la vida diaria. Las escalas de comportamientos creativos son los instrumentos más desarrollados por los investigadores interesados en el análisis de la creatividad cotidiana. Sin embargo, es preciso subrayar las importantes contribuciones que actualmente realizan los investigadores que desarrollan valoraciones momentáneas de la creatividad. Estas valoraciones suponen procesos dinámicos de análisis de acciones creativas en contextos naturales (Silvia et al. 2017).

Entre los instrumentos tradicionales y más utilizados para la valoración de comportamientos creativos, se destaca el Inventario de Comportamiento Creativo de Hocevar (1979). Este instrumento de autoinforme consta de 90 ítems a partir de los cuales se evalúan los logros en diferentes dominios: música, literatura, manualidades, artes, artesanías, performance y ciencias. El Cuestionario de Logros Creativos de Carson et al. (2005) también es un instrumento muy utilizado en el campo de la creatividad, evalúa actuaciones creativas en las dimensiones arte, música, danza, diseño, literatura, humor, inventos, descubrimientos cinéticos, teatro y cine y artes culinarias. Cada área incluye 7 ítems. El cuestionario evalúa cuántas veces los sujetos han logrado un rendimiento creativo. El Inventario Biográfico de Comportamientos Creativos de Batey (2007) incluye 34 actividades vinculadas a la creatividad cotidiana. Los participantes informan si han estado involucrados en estas actividades en los últimos 12 meses, este instrumento no evalúa la frecuencia de realización de la actividad. Kaufman (2012) creó la Escala de Dominios de Creatividad, que incluye 50 ítems que se refieren a logros creativos en cinco dominios: diario, académico, científico / mecánico, artístico y de actuación (incluye música y escritura). En Argentina, podemos mencionar a la Escala de Comportamiento Creativo (Aranguren & Irrazabal, 2012) diseñada con el propósito de evaluar desempeños creativos en los siguientes dominios: artes, manualidades, diseño, literatura, música, expresión corporal y negocios. Las personas deben responder la cantidad de veces que han participado de cada actividad considerando una escala Likert.

Desde el 2016, las publicaciones sobre instrumentos para medir la creatividad cotidiana han tomado perspectivas que integran diferentes dominios y manifestaciones creativas. Se destaca el Inventario de Actividades y Logros Creativos (ICAA) que consiste en una evaluación de las diferencias individuales en la creatividad de la vida cotidiana. El instrumento proporciona escalas independientes para la frecuencia de participación en la actividad creativa diaria y el nivel de logro creativo en ocho dominios creativos (Dietrich et al., 2018; Karwowski & Beghetto, 2018). El

ICAA es una medida de autoinforme que evalúa actividad creativa (CACT) y nivel de logro creativo (CACH). La escala CACT es conceptualmente similar al CBI (Hocevar, 1979) y pregunta con qué frecuencia se ha realizado una determinada actividad en los últimos 10 años. La escala CACH evalúa 11 niveles diferentes de logro creativo por dominio. Los participantes verifican qué nivel de logro aplica a ellos en un determinado dominio en una escala de 0 a 10. También interesa mencionar el estudio de Paek y Runco (2017) en el que se incorpora las escalas creatividad cotidiana y creatividad tecnológica a la Lista de Verificación de Actividades y Comportamientos Creativos (CAAC), el instrumento que permite comparar cantidad y calidad de actividades creativas en diferentes dominios. Los autores han reportado adecuados índices de confiabilidad del instrumento y diferencias según sexo.

En suma, los instrumentos construidos para valorar la creatividad cotidiana incluyen diferentes dominios e intentan valorar acciones creativas diversas. Algunos también valoran logros y reconocimientos en cada dominio. Es interesante señalar que los instrumentos actuales incluyen áreas referidas a las tecnologías considerando el impacto de las mismas en la vida cotidiana de las personas. Destacamos el valor de estos cuestionarios como herramientas útiles para analizar acciones y desempeños creativos; sin embargo, reconocemos limitaciones, como por ejemplos las señaladas por Silvia et al., (2017). Los autores consideran que este tipo de cuestionarios pone demasiado énfasis en la amplitud y en la frecuencia de participación, sin considerar el tipo de dedicación que las personas tienen por cada dominio. Esto quiere decir que los instrumentos no permiten captar si el desempeño es superficial o profundo en cada actividad. Asimismo, los autores señalan que este tipo de instrumento valora dominios tradicionales, sin incluir acciones creativas específicas de ciertos contextos o dominios excepcionales. A pesar de las debilidades señaladas, consideramos que las escalas de acciones creativas aportan datos relevantes para la comprensión de la creatividad cotidiana valorando el desempeño de las personas en diferentes áreas.

Racionalidad en el CAC42

Las acciones creativas son manifestaciones culturales que involucran interacciones con otras personas mediante la integración del pensamiento creativo y los comportamientos. Conviene recordar en ese sentido que las personas no son creativas en todas las acciones que realicen, sino más bien en unas pocas, y que sus manifestaciones creativas estarán siempre mediadas por el medio social en que están y por el contexto de los conocimientos y los niveles de desarrollo que hayan alcanzado. El concepto de acción creativa abarca dimensiones psicológicas, conductuales y culturales e integra cognición y conducta creativa en determinado contexto social (Glăveanu et al., 2019).

El CAC (Elisondo & Donolo, 2016) evalúa la frecuencia de participación en determinadas acciones creativas. El CAC evalúa actividades creativas de manera similar al CBI (Hocevar, 1979) y BICB (Batey, 2007) con respecto a la aplicación de diferentes dominios, con la ventaja adicional de incluir los dominios de *participación social* y *creatividad cotidiana*. Consideramos importante incluir estos dominios teniendo en cuenta la relevancia teórica y práctica de los estudios sobre creatividad cotidiana (Richards, 2010) y los enfoques socioculturales de los procesos creativos (Glăveanu, 2014). Los elementos del dominio *creatividad cotidiana* fueron construidos por considerar desarrollos teóricos e ítems de otros instrumentos CAQ (Carson et al., 2005), ICAA

(Dietrich et al., 2018), CAAC (Paek & Runco, 2017), y K-DOCS (Kaufman, 2012). El dominio *participación social* destaca acciones que involucran interacciones con otras personas e interés en la construcción de grupos e instituciones, estos ítems se basan en perspectivas socioculturales de la creatividad (Glăveanu, 2013; 2018). Las acciones creativas en *participación social* se orientan hacia el liderazgo y la transformación social.

La versión original del Cuestionario de Acciones Creativas (CAC) consta de 70 ítems que evalúan acciones creativas en 7 áreas: *literatura*, *artes plásticas y artesanías*, *música*, *ciencia y tecnología*, *participación social*, *creatividad cotidiana* y *expresión corporal* (Elisondo & Donolo, 2016). Cada dominio incluye 10 ítems que refieren a acciones creativas concretas como, por ejemplo: *ha pintado una obra*, *ha publicado un trabajo científico*, *actuó en teatro, cine o televisión*, etc. Las personas deben responder considerando las siguientes opciones: 1 (*nunca o casi nunca se ha hecho lo que se dice*), 2 (si *pocas veces lo ha hecho*, *2 o 3 veces*), 3 (si *lo ha hecho varias veces, 4 o 5 veces*), 4 (si *lo hizo frecuentemente, 6 o 7 veces*) y 5 (si *en la mayoría de las veces que tuvo oportunidad lo hizo*).

En el estudio de la versión original del CAC (Elisondo & Donolo, 2016) hemos observado una adecuada consistencia interna entre los ítems de cada área a través de la prueba alfa de Cronbach: $\alpha = .849$ en *creatividad cotidiana*; $\alpha = .823$ en *música*; $\alpha = .823$ en *artes plásticas y artesanías*; $\alpha = .799$ en *expresión corporal*; $\alpha = .84$ en *participación social*; $\alpha = .71$ en *literatura* y $\alpha = .67$ en *ciencia y tecnología*. Además, hallamos correlaciones positivas y superiores a .50 entre cada ítem y su respectiva área de conocimiento, y correlaciones bajas entre ítems y otras áreas de conocimiento. Respecto de las relaciones con otros instrumentos, encontramos una correlación significativa de $r = .60$ entre el CAC y el BICB. Los resultados también indicaron diferencias significativas en el CAC según sexo, escolaridad y participación en actividades de ocio.

Método

Se presenta un estudio instrumental (Montero & León, 2007), es decir, una investigación orientada al desarrollo de pruebas y el análisis de propiedades psicométricas de los instrumentos.

Participantes

La muestra se conformó de manera no probabilística y por conveniencia, se procuró incluir personas de diferentes edades, niveles de escolaridad y lugares de procedencia. Participaron del estudio 1509 personas. El 68% de los participantes expresaron ser mujeres y el resto se definieron como hombres. Los participantes tienen entre 18 y 94 años, con contribuciones etarias de este tenor: 18-30 años (58%), 31-60 años (28%) y más de 60 años (14%). Se aprecia que la mayor contribución porcentual es de jóvenes. Los participantes residen en diferentes provincias de Argentina: son principalmente de Córdoba (80%), le sigue en orden decreciente Buenos Aires (12%); San Luis y La Pampa (2%); Salta, Mendoza, Misiones (1%) y el resto del 1% para las demás provincias. Todos los participantes hablan español y manifestaron pertenecer a un nivel socioeconómico medio. En la muestra incluimos personas que han cursado estudios hasta nivel secundario completo o incompleto (53%) o han iniciado o completado el nivel superior (47%).

Instrumentos

En la presente investigación se utilizaron cuatro instrumentos: CAC, BICB, cuestionario sociodemográfico y cuestionario de ocio. Todos los participantes respondieron a la versión original del CAC y al cuestionario sociodemográfico (sexo, edad, escolaridad y procedencia geográfica). El cuestionario sobre ocio fue respondido por 488 participantes. Los ítems de este cuestionario indagan respecto del tipo de actividades desarrolladas por los participantes y nivel de compromiso en dichas acciones (frecuencia, duración y tipo de participación). Doscientos noventa y nueve participantes (N = 299) respondieron al BICB (Batey, 2007), dicho inventario evalúa la creatividad cotidiana considerando 34 actividades respecto de las cuales las personas deben indicar si participaron activamente en ellas en los últimos 12 meses. El inventario permite dos tipos de respuestas: afirmativa o negativa y arroja una puntuación total de comportamiento creativo. En estudios anteriores, el BICB ha demostrado tener la consistencia adecuada interna (α = .78) y correlaciones significativas con otras medidas de la creatividad (Batey, Furnham, & Safiullina, 2010). El BICB fue traducido al español por una traductora pública nacional. La versión traducida fue revisada por cinco expertos de habla hispana con conocimientos avanzados de idioma inglés.

Procedimientos y análisis

Los instrumentos fueron administrados online. Contamos con el consentimiento informado de los participantes para la realización de la investigación y la publicación de resultados, preservando la confidencialidad de los datos. Se realizaron diferentes análisis estadísticos con los programas SPSS20 (IBM, 2011), AMOS (Arbuckle, 2014), JAMOVI (Jamovi Project, 2021) y JMETRIK (Jmetrik Group, 2014). Se realizaron análisis de frecuencias, media, desviaciones estándar, diferencia de media, coeficiente omega de Mc Donald, correlaciones de Pearson, análisis factorial confirmatorio y estudio diferencial de ítems.

Resultados

Construcción del CAC42

Para la construcción del CAC42 se analizaron correlaciones entre ítems y dominios, se eliminaron aquellos que correlacionaban con menor intensidad: 28 ítems (4 por cada dominio). Asimismo, se verificó la capacidad discriminativa de cada ítem considerando dos grupos, el primero conformado por casos de puntajes altos en cada área (por encima del percentil 75) y el segundo con puntajes bajos en cada área (por debajo del percentil 25). En todos los ítems seleccionados para la versión CAC42 se observaron diferencias significativas entre los dos grupos. También se calculó el índice de discriminación para cada ítem de cada área. Los ítems seleccionados para el CAC42 tienen índices de discriminación por encima de .35. En la Lista 1 se indican los ítems eliminados de la versión original del CAC.

Análisis de la escala

La Tabla 1 muestra estadísticas descriptivas para cada dominio y puntaje total en la versión abreviada. Al igual que en la versión original (Elisondo & Donolo 2016), se observa que los participantes realizan con más frecuencia acciones creativas en el dominio *creatividad cotidiana* y con menos frecuencia actividades referidas a *ciencia y tecnología*. Se calculó el coeficiente omega de McDonald con un intervalo de confian-

Lista 1: Ítems de la versión original del CAC, con detalle de aquellos eliminados para la construcción del CAC42.

1. Encontró soluciones a sus problemas mirándolos desde diferentes puntos de vista.
2. Ejecuta un instrumento musical *
3. Ha pintado una obra
4. Actuó en teatro, cine o televisión
5. Participó en una ONG, partido político o comunidad religiosa *
6. Ha publicado un trabajo literario
7. Ha diseñado una página web *
8. Encontró diferentes formas de entretener a un niño.
9. Escribió música para instrumentos
10. Realizó artesanías de metal, madera, plástico, vidrio, cuero, cerámica.
11. Participó en una asociación, club u organización de actuación o danza
12. Ha organizado eventos comunitarios y sociales
13. Ha recibido un premio por su trabajo literario
14. Investigó utilizando teorías, métodos e instrumentos no convencionales *
15. Encontró una manera nueva y eficiente de ordenar sus objetos personales. *
16. Ganó un premio por su habilidad musical
17. Ha inventado un objeto o modelo original *
18. Bailó en una compañía de danza
19. Ha liderado un grupo o actividad social
20. Trabajó como editor *
21. Ha creado un programa de computación
22. Descubrió diferentes formas de “llegar a fin de mes “ *
23. Grabó un disco *
24. Diseñó y realizó prendas de vestir *
25. Dirigió una obra de teatro *
26. Ha creado un nuevo sistema de organización en una institución social
27. Fundó una revista o publicación periódica *
28. Ha publicado un trabajo científico.
29. Ayudó a otros a afrontar situaciones difíciles
30. Participó en una competencia musical
31. Ha presentado sus pinturas, esculturas o fotografías en eventos artísticos
32. Recibió un premio por su actuación
33. Ha recibido un premio por su participación en ONGs. *
34. Participó en una organización de escritores
35. Ha ganado un premio en un evento científico o tecnológico
36. Convenció a alguien de hacer algo.
37. Protagonizó un recital
38. Ha recibido un premio por sus pinturas, esculturas o fotografías *
39. Realizó una coreografía original
40. Ha creado una ONG, agrupación o institución social
41. Escribió un artículo para un diario o una revista
42. Participó en una asociación u organización científica o tecnológica
43. Ha buscado diferentes alternativas a problemas sociales *
44. Compuso música con recursos informáticos *
45. Realizó decoraciones de espacios interiores o exteriores *
46. Ha producido una obra de teatro, cine, televisión o radio *
47. Ha dado discursos y participado en debates públicos. *
48. Ha escrito un trabajo literario extenso
49. Diseñó un experimento para explicar algo
50. Reorganizó su vida integrando necesidades personales, familiares y laborales.
51. Fue miembro de un grupo musical
52. Creó títeres o marionetas *
53. Actuó en un ballet, show o competencia dramática
54. Ha creado nuevos proyectos comunitarios
55. Escribió una obra de teatro *
56. Escribió letras ingeniosas o humorísticas *
57. Planeó un viaje en el que tuvo en cuenta los intereses de todos los viajeros. *
58. Creó un instrumento musical *
59. Participó en exposiciones de artesanías, vestimentas, recetas o decoraciones. *
60. Tuvo un papel en una producción dramática *
61. Ha generado espacios que promueven la participación social
62. Le han otorgado una beca de formación o investigación científica
63. Descubrió nuevas formas de ayudar a la gente.
64. Compuso una música original que ha sido presentada públicamente
65. Realizó joyas o accesorios (aros, colgantes, etc.)
66. Ganó un premio por sus artesanías *
67. Realizó un arreglo floral original
68. Ha realizado una escultura.
69. Creó tarjetas de salutación o invitación a eventos *
70. Ha escrito un trabajo literario corto

* Ítem eliminado para la construcción de la versión abreviada.

Tabla 1

Estadística descriptiva para los siete dominios y el puntaje total del CAC42 para 1509 casos.

Dominios	M	DE	Mín	Máx	Asimetría	Curtosis
Artes Plásticas y Artesanías	9.58	3.62	6	28	1.52	2.89
Creatividad cotidiana	19.46	5.42	6	30	-0.04	-0.66
Música	6.54	1.55	6	20	4.32	22.78
Artes Escénicas	8.81	4.03	6	30	2.04	4.21
Participación Social	8.53	3.34	6	27	2.01	5.50
Literatura	7.10	1.62	6	18	1.97	4.77
Ciencia y Tecnología	6.46	1.02	6	16	3.21	13.85
Total CAC42	66.50	11.99	42	119	0.645	0.21

za de .95. En la versión abreviada, los elementos de cinco dominios muestran una consistencia interna satisfactoria (ω > .70), *artesanías y artes plásticas*: ω = .71; *creatividad cotidiana* ω = .79; *música* ω = .75 (CAC42); *expresión corporal* ω = .81; *participación social* ω = .76. En cambio, los índices del coeficiente omega fueron más bajos en los dominios *literatura* ω = .62 y *ciencia y tecnología* ω = .61. En la versión original (Elisondo & Donolo, 2016) también se observaron menores índices de confiabilidad en *ciencia y tecnología* y *literatura*. El análisis factorial confirmatorio indicó un adecuado ajuste del modelo del CAC42. Los índices de ajuste ($\chi^2_{(798)}$ = 2487, p < .001), índice de ajuste comparativo [CFI] = .990, índice de bondad de ajuste [GFI] = .999, error cuadrático medio de aproximación [RMSEA] = .037), sugieren que el modelo postulado se ajusta razonablemente a los datos. Los ítems muestran adecuadas cargas factoriales (.45 a .99) en su respectiva variable latente (ver Tabla 2). Los análisis también indicaron correlaciones significativas entre algunas variables latentes.

Validez

Grupos de contraste. A partir de las respuestas al cuestionario de ocio, se definieron dos grupos: personas que participan activamente en alguna o varias propuestas artísticas (n = 242) y personas que no participan (n = 246). Las personas que participan de manera comprometida han manifestado hacerlo en contextos particulares (talleres, instituciones educativas, organizaciones no gubernamentales, etc.) dedicando más de 5 horas semanales a la actividad, en los últimos dos años. Los participantes manifestaron desempeñarse de manera comprometida en alguna de las siguientes actividades: teatro, artes plásticas, artesanías, artes culinarias, música, expresión corporal y danza.

Se evaluó el funcionamiento diferencial de todos los ítems según participación en actividades de ocio. Se observó funcionamiento diferencial en los ítems 11 y 63 del CAC42 según *ocio*. Se realizaron estudios de diferencia de media entre personas que participan de manera comprometida en actividades de ocio y quienes no lo hacen. Los resultados indican diferencias significativas en las áreas *música*, *artes plásticas y artesanías*, *artes escénicas* y *participación social*, las personas que participan en este tipo de actividades obtuvieron puntajes medios significativamente superiores (ver Tabla 3).

Validez concurrente con el BICB. Los análisis

Tabla 2

Análisis factorial confirmatorio: cargas factoriales para cada ítem del CAC42.

	Creatividad cotidiana	Música	Artes plásticas y Artesanías	Artes escénicas	Participación social	Literatura	Ciencia y Tecnología
V1	.57						
V8	.49						
V29	.76						
V36	.68						
V50	.65						
V63	.66						
V9		.44					
V16		.98					
V30		.57					
V37		.71					
V51		.71					
V64		.89					
V3			.58				
V10			.56				
V31			.60				
V65			.63				
V67			.47				
V68			.59				
V18				.72			
V11				.72			
V4				.59			
V32				.60			
V39				.73			
V53				.76			
V12					.73		
V19					.68		
V26					.58		
V40					.57		
V54					.76		
V61					.84		
V6						.59	
V13						.46	
V34						.77	
V41						.64	
V48						.56	
VA70						.45	
V21							.45
V28							.75
V35							.60
V42							.95
V49							.45
V62							.70
Música	.39						
Artes plásticas y Artesanías	.32	.74					
Artes escénicas	.37	.95	.67				
Participación social	.44	.80	.52	.62			
Literatura	.38	.99	.74	.87	.78		
Ciencia y Tecnología	.35	.99	.68	.83	.76	.93	

$X^2_{(798)} = 2{,}487$ $p < .001$ [CFI] = .99 [GFI] = .99 [RMSEA] = .03

Tabla 3

Media, desviación estándar y prueba *t* en CAC42 según participación en actividades de ocio.

	Participa n = 242		No participa n = 246				95% CI		
	M	DE	M	DE	$t_{(486)}$	*p*	LL	UL	Cohen´s *d*
Artes Plásticas y Artesanías	9.87	3.70	9.04	3.10	2.78	.00	-1.43	-0.21	.24
Creatividad Cotidiana	20.14	5.12	19.78	4.69	0.81	.41	-0.51	1.23	.07
Ciencia y Tecnología	6.80	1.72	6.76	1.52	0.30	.76	-0.24	0.33	.04
Música	7.22	3.25	6.62	2.14	2.41	.01	-1.09	-0.11	.22
Literatura	7.09	2.00	7.07	1.99	0.01	.98	-0.34	0.35	.01
Artes Escénicas	9.46	4.10	8.57	3.67	2.52	.01	-1.58	-0.19	.22
Participación Social	9.52	5.56	8.32	3.42	3.42	.00	-1.88	-0.51	.25
Total CAC	69.72	13.80	66.61	12.29	2.62	.00	-5.42	-0.78	.23

indican correlaciones positivas y moderadas en el CAC versión abreviada y el BICB (r = .582, p < .001). Resultados similares se observaron en el estudio con la versión original del CAC (Elisondo & Donolo, 2016).

Discusión

Los resultados hallados dan cuenta de las propiedades psicométricas del CAC42 y muestran similitudes con los datos de un estudio anterior con la versión original del CAC (Elisondo & Donolo 2016). En el presente estudio observamos índices de consistencia satisfactorios y un adecuado ajuste del modelo propuesto que incluye siete dominios creativos. Asimismo, al igual que en el estudio preliminar (Elisondo & Donolo, 2016), encontramos diferencias según la participación en actividades artísticas, resultados que aportan evidencias de validez para el CAC42. También hallamos evidencias de validez convergente a partir del estudio correlacional con el BICB.

Los análisis indican diferencias en acciones creativas según participación comprometida en actividades artísticas. Las personas que participan en estas actividades obtuvieron puntajes significativamente superiores en los dominios artísticos y en participación social. No se observan diferencias en *creatividad cotidiana* y *ciencia y tecnología*. Estos resultados se vinculan con los hallados por Aranguren e Irrazabal (2012): diferencias estadísticamente significativas en el puntaje total de Escala de Comportamiento Creativo y en las subescalas de *artes y diseño*, *literatura y música* y *expresión corporal* entre los participantes que habían desarrollado alguna actividad artística y aquellos que no habían realizado ninguna actividad. Los resultados hallados en la presente investigación también se vinculan con datos de otros estudios que indican relaciones entre participación en actividades de ocio y creatividad (Hegarty, 2009; Wolfradt & Pretz, 2001). En el presente estudio, la participación comprometida en actividades artísticas se relaciona con el desarrollo de acciones creativas vinculadas a la música, las artes plásticas, las artesanías y las artes escénicas. Resulta interesante también que las actividades de ocio generan oportunidades para el desarrollo de acciones creativas de participación

social, en tanto potencia interacciones con otras personas y procesos de construcción de grupos y organizaciones.

Asimismo, los resultados aportan evidencias en el proceso de validación del CAC42, en futuros estudios es relevante triangular datos con otras técnicas que permitan comprender la complejidad de las acciones creativas en contextos cotidianos. Los estudios que se basan en métodos de muestreo de experiencia (Silvia et al., 2014) pueden hacer contribuciones interesantes a las investigaciones de la creatividad y aportar datos respecto de dominios y campos donde se desarrollan acciones creativas. Incluir otros dominios como los negocios y los deportes es una interesante línea para investigaciones futuras. Considerando las críticas de Silvia et al. (2017) a los cuestionarios de comportamientos creativos, en otros estudios podría analizarse el tipo de compromiso y dedicación de las personas a las acciones creativas y asignar puntajes adicionales en el CAC42. Entre las limitaciones del estudio señalamos: el uso de un instrumento no validado en español (BICB), la no inclusión de otros instrumentos que miden variables vinculadas a la creatividad, como la personalidad y la inteligencia, y el trabajo con muestras no probabilísticas. El estudio también presenta limitaciones en cuanto a los análisis realizados, se propone en futuros estudios presentar modelos más complejos y considerar las diferencias observadas entre los diferentes grupos.

Sin embargo, destacamos el valor del CAC42 como instrumento que ofrece indicadores generales respecto de acciones creativas en diferentes dominios. También es relevante señalar que el CAC42 se sustenta en perspectivas actuales de la creatividad cotidiana y en paradigmas socioculturales integrando contribuciones de estos enfoques conceptuales. Entre las fortalezas del presente estudio, también se encuentran la amplitud de la muestra, que incluyó a 1509 participantes, y la presencia de dos dominios novedosos (creatividad diaria y participación social). En resumen, el CAC42 es una herramienta útil para evaluar acciones creativas en diversos dominios que se basa en las teorías actuales de la creatividad. El CAC42 puede ser una herramienta valiosa de evaluación para la toma de decisiones en contextos científicos, educativos y laborales.

Referencias

Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 153-158. doi: 10.1037/aca0000231

Adams-Price, C. E., & Steinman, B. A. (2007). Crafts and generative expression: A qualitative study of the meaning of creativity in women who make jewelry in midlife. *The International Journal of Aging and Human Development, 65*(4), 315-333. doi: 10.2190/AG.65.4.c

Aranguren, M., & Irrazabal, N. (2012). Diseño de una escala para la evaluación del comportamiento creativo. *Ciencias Psicológicas, 6*(1), 29-41. doi: 10.22235/cp.v6i1.60

Arbuckle, J. L. (2014). Amos (Versión 23.0). [Software de cómputo]. Chicago: IBM SPSS.

Barbot, B., Hass, R. W., & Reiter-Palmon, R. (2019). Creativity assessment in psychological research: (Re)setting the standards. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 233-240. doi: 10.1037/aca0000233

Baer, J. (2012). Domain specificity and the limits of creativity theory. *The Journal of Creative Behavior, 46*(1), 16-29. doi: 10.1002/jocb.002

Batey, M. (2007). *A psychometric investigation of everyday creativity* (Tesis de doctorado). University College, London.

Batey, M., Furnham, A., & Safiullina, X. (2010). Intelligence, general knowledge and personality as predictors

of creativity. *Learning and Individual Differences, 20*(5), 532-535. doi: 10.1016/j.lindif.2010.04.008

Benedek, M., Bruckdorfer, R., & Jauk, E. (2019). Motives for Creativity: Exploring the what and why of everyday creativity. *The Journal of Creative Behavior, 54*(3), 610-625. doi: 10.1002/jocb.396

Benedek, M., Christensen, A. P., Fink, A., & Beaty, R. E. (2019). Creativity assessment in neuroscience research. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 218-226. doi: 10.1037/aca0000215

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the Creative Achievement Questionnaire. *Creativity Research Journal, 17*(1), 37-50. doi: 10.1207/s15326934crj1701_4

Conner, T. S., DeYoung, C. G., & Silvia, P. J. (2018). Everyday creative activity as a path to flourishing. *The Journal of Positive Psychology, 13*(2), 181-189. doi: 10.1080/17439760.2016.1257049

Conner, T. S., & Silvia, P. J. (2015). Creative days: A daily diary study of emotion, personality, and everyday creativity. *Psychology of Aesthetics, Creativity, and the Arts, 9*(4), 463-470. doi: 10.1037/aca0000022

Diedrich, J., Jauk, E., Silvia, P. J., Gredlein, J. M., Neubauer, A. C., & Benedek, M. (2018). Assessment of real-life creativity: The Inventory of Creative Activities and Achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts, 12*(3), 304-316. doi: 10.1037/aca0000137

Elisondo, R., & Donolo, D. (2016). Construcción y análisis de las propiedades psicométricas del Cuestionario de Acciones creativas en población argentina. *PSIENCIA. Revista Latinoamericana de Ciencia Psicológica, 8*(1), 1-21. Recuperado de https://www.redalyc.org/journal/3331/333145838006

Elisondo, R., & Donolo, D. (2018). Contextos y creatividad. Variables sociodemográficas y datos normativos en el Test CREA. *Revista Evaluar, 18*(3), 14-29. doi: 10.35670/1667-4545.v18.n3.22202

Elisondo, R. C., Donolo, D. S., & Corbalan-Berná, F. J. (2009). Evaluación de la Creatividad ¿Relaciones con inteligencia y personalidad? *Revista Iberoamericana de Diagnóstico y Evaluación -e Avaliação Psicológica, 2*(28), 67-79. Recuperado de https://www.aidep.org/es/numeros-publicados

Elisondo, R. C., & Vargas, A. (2019). Women's everyday creative activities: A qualitative study. *Creativity. Theories-Research-Applications, 6*(1), 91-111. doi: 10.1515/ctra-2019-0006

Gandolfo, E., & Grace, M. (2010) Women doing it forever: The everyday creativity of women craftmakers. *Australian and New Zealand Journal of Art Therapy, 5*(1), 29-44. Recuperado de https://www.jocat-online.org

Garcês, S., Pocinho, M., Neves de Jesus, S. N., Viseu, J., Imaginário, S., & Muglia-Wechsler, S. M. (2015). Estudo de Validação da Escala de Personalidade Criativa. *Revista Iberoamericana de Diagnóstico y Evaluación - e Avaliação Psicológica, 2*(40), 17-24. Recuperado de https://www.aidep.org/es/numeros-publicados

Genoe, M. R., & Liechty, T. (2017) Meanings of participation in a leisure arts pottery programme. *World Leisure Journal, 59*(2), 91-104. doi: 10.1080/16078055.2016.1212733

Glăveanu, V. P. (2013). Rewriting the language of creativity: The five A's framework. *Review of General Psychology, 17*(1), 69-81. doi: 10.1037/a0029528

Glăveanu, V. P. (2014). The psychology of creativity: A critical reading. *Creativity: Theories-Research-Applications, 1*(1), 10-32. doi: 10.15290/ctra.2014.01.01.02

Glăveanu, V. P. (2018). Educating which creativity? *Thinking Skills and Creativity, 27*, 25-32. doi: 10.1016/j.tsc.2017.11.006

Glăveanu, V. P., Hanchett-Hanson, M., Baer, J., Barbot, B., Clapp, E. P., Corazza, G. E., ... & Stenberg, R. J. (2019). Advancing creativity theory and research: A socio-cultural manifesto. *The Journal of Creative Behavior, 54*(3), 741-745. doi: 10.1002/jocb.395

Hegarty, C. B. (2009). The value and meaning of creative leisure. *Psychology of Aesthetics, Creativity, and the*

Arts, 3(1), 10-13. doi: 10.1037/a0014879

Hocevar, D. (1979). The development of the Creative Behavior Inventory. *Annual Meeting of the Rocky Mountain Psychological Association.* Las Vegas, USA. Recuperado de https://eric.ed.gov

IBM. (2011). SPSS Statistics for Windows (Version 20.0). [Software de cómputo]. Armonk, NY: IBM Corp.

Jamovi project. (2021). Jamovi (Version 1.6). [Software de cómputo]. Recuperado de https://www.jamovi.org

Jmetrik Group. (2014) Jmetrik (Version 4.4.1). [Software de cómputo]. Recuperado de https://itemanalysis.com/jmetrik-download

Karwowski, M., & Beghetto, R. A. (2018). Creative behavior as agentic action. *Psychology of Aesthetics, Creativity, and the Arts, 13*(4), 402-415. doi: 10.1037/aca0000190

Kaufman, J. C. (2006). Self-reported differences in creativity by gender and ethnicity. *Applied Cognitive Psychology, 20*(8), 1065-1082. doi: 10.1002/acp.1255

Kaufman, J. C. (2012). Counting the muses: Development of the Kaufman Domains of Creativity Scale (K-DOCS). *Psychology of Aesthetics, Creativity, and the Arts, 6*(4), 298-308. doi: 10.1037/a0029751

McCabe, M., & de Waal-Malefyt, T. (2015). Creativity and cooking: Motherhood, agency and social change in everyday life. *Journal of Consumer Culture, 15*(1), 48-65. doi: 10.1177/1469540513493202

Modrzejewska-Świgulska, M. (2018). Professional Competences. Reconstruction of the opinions of Polish female directors. *Creativity. Theories-Research-Applications, 5*(1), 72-83. doi: 10.1515/ctra-2018-0005

Montero, I., & León, O. G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology, 7*(3), 847-862. Recuperado de https://aepc.es/ijchp/busca.php

Paek, S. H., & Runco, M. A. (2017). Dealing with the criterion problem by measuring the quality and quantity of creative activity and accomplishment. *Creativity Research Journal, 29*(2), 167-173. doi: 10.1080/10400419.2017.1304078

Richards, R. (2010). Everyday creativity. Process and way of life - Four key issues. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge Handbook of Creativity* (pp. 189-215). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511763205.013

Silvia, P. J., Beaty, R. E., Nusbaum, E. C., Eddington, K. M., Levin-Aspenson, H., & Kwapil, T. R. (2014). Everyday creativity in daily life: An experience-sampling study of "little c" creativity. *Psychology of Aesthetics, Creativity, and the Arts, 8*(2), 183-188. doi: 10.1037/a0035722

Silvia, P. J., Cotter, K. N., & Christensen, A. P. (2017). The creative self in context: Experience sampling and the ecology of everyday creativity. In M. Karwowski & J. C. Kaufman (Eds.), *The creative self: Effect of beliefs, self-efficacy, mindset, and identity* (pp. 275-288). Cambridge, Massachusetts: Elsevier Academic Press. doi: 10.1016/B978-0-12-809790-8.00015-7

Sternberg, R. J. (2009). Domain-generality versus domain-specificity of creativity. In P. Meusburger, J. Funke & E. Wunder (Eds.), *Milieus of Creativity* (pp. 25-38). Dordrecht: Springer. doi: 10.1007/978-1-4020-9877-2_3

Wolfradt, U., & Pretz, J. E. (2001). Individual differences in creativity: Personality, story writing and hobbies. *European Journal of Personality, 15*(4), 297-310. doi: 10.1002/per.409